AD-A087 111 COLORADO STATE UNIV FORT COLLINS DEPT OF STATISTICS F/G 12/2
TESTING FOR A CHANGE POINT.(U)
APR 80 N J FARDEN N00014-75-C-0518
UNCLASSIFIED TR-19 NL

1 OF 2
AD A
087111

Testing for a Change Point

by

Nancy Jean Farden

ONR Technical Report #19

April 1980

Prepared for the Office of Naval Research
under Contract N00014-75-C-0518

L. L. Scharf, Principal Investigator

Testing for a Change Point[1]

by

Nancy Jean Farden[2]

[1]This work partially supported by the Office of Naval Research, Statistics and Probability Branch, Arlington, VA, under Contract N00014-75-C-0518.

[2]Department of Statistics, Colorado State University, Fort Collins, Colorado 80523.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER Technical Report #19	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Testing for a Change Point		5. TYPE OF REPORT & PERIOD COVERED Technical Report
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Nancy Jean Farden		8. CONTRACT OR GRANT NUMBER(s) N00014-75-C-0518
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics Colorado State University Fort Collins, Colorado 80523		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS		12. REPORT DATE April 1980
		13. NUMBER OF PAGES 180
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) Office of Naval Research Statistics and Probability Branch Arlington, VA		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

Unlimited distribution

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Bernoulli trials, hypothesis testing, change points, experimental design,

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

A nonparametric procedure based on differences of averages is proposed to test for a change in the location parameter of the distribution for a sequence of independent random variables. We assume that at most one change may have occurred and that the point of change is an unknown parameter; the initial value of the location parameter is assumed known.

The emphasis is on obtaining a method of testing for a change point for small sample sizes which is easy to use and which will detect a change

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Abstract (con't)

quickly.

Exact critical regions are tabulated for $n \leq 20$; the power of the tests and properties of the estimator for the change point are studied. Appropriate design of experiments for detecting change points are suggested.

ABSTRACT OF DISSERTATION

TESTING FOR A CHANGE POINT

A nonparametric procedure based on differences of averages is proposed to test for a change in the location parameter of the distribution for a sequence of independent random variables. We assume that at most one change may have occurred and that the point of change is an unknown parameter; the initial value of the location parameter is assumed known.

The emphasis is on obtaining a method of testing for a change point for small sample sizes which is easy to use and which will detect a change quickly.

Exact critical regions are tabulated for $n \leq 20$; the power of the tests and properties of the estimator for the change point are studied. Appropriate design of experiments for detecting change points are suggested.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

LIST OF TABLES AND FIGURES

LIST OF TABLES AND FIGURES (Continued)

LIST OF TABLES AND FIGURES (Continued)

CHAPTER I

INTRODUCTION

1.1 Motivation

Suppose that a process is sampled at consecutive points and it is known that at least the first observations satisfy certain distributional properties, but after some unknown point the remaining observations may satisfy other distributional properties. It is of interest to determine whether such a change has occurred, and if so, to estimate the point of change and the current properties of the process. This situation may arise in a number of applications. For example, suppose specifications for the manufacturing of batteries allows for no more than a certain percentage of defectives to be produced; a method for determining whether the percentage has exceeded this value is needed. If the manufacturer decides that product quality has deteriorated, he may want to estimate when the percentage of defectives increased so that these batteries are not sold. As another example, consider the sales record of wool fabric; if sales increase production may need to be stepped up to keep the fabric available, or if sales decrease the production may need to be slowed down. In this instance, not only is the estimate of when a change in sales occurred of interest, but also an estimate of the current demand for the product. A third

example deals with the medical experimentation of a new drug for treatment of pain. If the data indicate that the drug is effective, the researcher may want to estimate the reaction time of the drug as well as a measurement of the patient's improvement for the dosage prescribed.

1.2 The Statement of the Problem

The following specific case is investigated to study processes for which a change in the model is suspected. Let $X_1, \cdots, X_n$ be independent Bernoulli random variables taking on the values 1 or -1 with $P(X_i=1) = 1 - P(X_i=-1) = p$ for $i = 1, \cdots, \rho$ and $P(X_i=1) = 1 - P(X_i=-1) = q$ for $i = \rho+1, \cdots, n$, where ρ is an unknown parameter, $1 \leq \rho \leq n$, and p and q are distinct Bernoulli parameters, where p is assumed known and q may be unknown. If $\rho < n$ we say that a change in the model has occurred. ρ is called the change point, that is, at step $\rho + 1$ the model changes.

We are interested in testing whether a change in the model has occurred. The hypotheses considered are

$$H_0:\ \rho=n, p \ \text{ vs. } \ H_1:\ \rho<n, p, q, p\neq q \ \text{ for the two-sided test with } p\neq q, \tag{1.2.1}$$

$$H_0:\ \rho=n, p \ \text{ vs. } \ H_1^+:\ \rho<n, p, q, p>q \ \text{ for the one-sided test with } p>q,$$

and $$H_0:\ \rho=n, p \ \text{ vs. } \ H_1^-:\ \rho<n, p, q, p<q \ \text{ for the one-sided test with } p<q.$$

We are furthermore interested in the situation where n is small, that is, $n \leq 20$. The aim is to obtain a test for small n which is easy to use and which will detect a change quickly.

The statistics we propose for testing for a change in the model are

$$S_n = S_n(X_1, \cdots, X_n) = \sup_r \left| \frac{1}{r} \sum_{i=1}^{r} X_i - \frac{1}{n-r} \sum_{i=r+1}^{n} X_i \right|$$

1.2.2 for H_0 vs. H_1,

$$S_n^+ = S_n^+(X_1, \cdots, X_n) = \sup_r \left\{ \frac{1}{r} \sum_{i=1}^{r} X_i - \frac{1}{n-r} \sum_{i=r+1}^{n} X_i \right\}$$

for H_0 vs. H_1^+

and $$S_n^- = S_n^-(X_1, \cdots, X_n) = \inf_r \left\{ \frac{1}{r} \sum_{i=1}^{r} X_i - \frac{1}{n-r} \sum_{i=r+1}^{n} X_i \right\}$$

for H_0 vs. H_1^-,

where $\frac{1}{r} \sum_{i=1}^{r} X_i$ and $\frac{1}{n-r} \sum_{i=r+1}^{n} X_i$ are taken to be zero when $r = 0$ and $r = n$, respectively. One intuitively expects S_n and S_n^+ to be "large" and S_n^- to be "small" when H_1, H_1^+ and H_1^-, respectively, are true. Note that $S_n \epsilon [0, 2]$, $S_n^+ \epsilon [0, 2]$ and $S_n^- \epsilon [-2, 0]$. Clearly $S_n \geq 0$; $S_n^+ \geq 0$ since $0 \leq \sup \left\{ \frac{1}{n} \sum_{i=1}^{n} X_i, -\frac{1}{n} \sum_{i=1}^{n} X_i \right\}$

$$= \sup_{r=0,\, n} \left\{ \frac{1}{r} \sum_{i=1}^{r} X_i - \frac{1}{n-r} \sum_{i=r+1}^{n} X_i \right\} \leq \sup_r \left\{ \frac{1}{r} \sum_{i=1}^{r} X_i - \frac{1}{n-r} \sum_{i=r+1}^{n} X_i \right\};$$

similarly $S_n^- \leq 0$. $S_n \leq 2$, $S_n^+ \leq 2$ and $S_n^- \geq -2$ since $\left| \frac{1}{r} \sum_{i=1}^{r} X_i \right| \leq 1$ and $\left| \frac{1}{n-r} \sum_{i=r+1}^{n} X_i \right| \leq 1$ for all r. Hence by "large" we mean "near 2" and by "small" we mean "near -2".

Define by R_n, R_n^+ and R_n^- the random variables corresponding to the points where the values of S_n, S_n^+ and S_n^-, respectively, are attained. For obvious reasons we will refer to R_n, R_n^+ and R_n^- as the change point random variables. To insure that these are bonafide random variables, they must be defined uniquely. Suppose that for the sequence $X_1, \cdots, X_n$, $S_n(X_1, \cdots, X_n) = \sup_r \left| \frac{1}{r} \sum_{i=1}^{r} X_i - \frac{1}{n-r} \sum_{i=r+1}^{n} X_i \right|$ occurs at $j = j_1, \cdots, j_\ell$. Denote by J_n the mapping $J_n(X_1, \cdots, X_n) = j_1, \cdots, j_\ell$. Randomize by defining the function U so that $P(U \circ J_n(X_1, \cdots, X_n) = j_i) = \frac{1}{\ell}$, $i = 1, \cdots, \ell$. Define $R_n(X_1, \cdots, X_n) = U \circ J_n(X_1, \cdots, X_n)$. Similarly randomize to determine R_n^+ and R_n^- uniquely. Fortunately for most sequences, the test statistics occur at a unique point; also for these sequences for which the supremum does not occur uniquely, ℓ is rarely larger than 2.

The change point random variables are of interest for the following reason. When the alternative hypothesis is true, we expect that the point r at which the statistic is attained should be "close" to the change point ρ. We will in fact estimate ρ to be this point in the data sequence; denote the estimator by $\hat{\rho}$. (Nonuniqueness difficulties will be resolved by randomizing.) When q is unknown, we use the last

$n-\hat{\rho}$ observations to estimate it; $\hat{q} = \frac{1}{n-\hat{\rho}} \sum_{i=\hat{\rho}+1}^{n} (\frac{X_i+1}{2})$ is an obvious estimator to consider.

1.3 Extensions

The statistics we have proposed for testing for a change in the model will be studied only for the case where the X_i's are 1 or -1 Bernoulli random variables, but the test can be used for more general situations as shown below.

Note that the test statistics are invariant under scale transformations and independent of location transformations. S_n is invariant for the scale transformation and independent of the location transformation $a X_i + b$ where $a \neq 0$; S_n^+ and S_n^- are invariant for the scale transformation and independent of the location transformation $a X_i + b$ where $a > 0$. Thus if $Y_1, \cdots, Y_n$ are independent Bernoulli random variables taking on the values c and d with $P(Y_i=c) = 1 - P(Y_i=d) = p$ for $i = 1, \cdots, \rho$ and $P(Y_i=c) = 1 - P(Y_i=d) = q$ for $i = \rho+1, \cdots, n$ we can test whether or not a change in the model has occurred; in particular for the two-sided test with $p \neq q$, choose $a \neq 0$ and b such that $Y_i = a X_i + b$, $i = 1, \cdots, n$, where $P(X_i=1) = 1 - P(X_i=-1) = p$ for $i = 1, \cdots, \rho$ and $P(X_i=1) = 1 - P(X_i=-1) = q$ for $i = \rho+1, \cdots, n$. We have

$$S_n(Y_1, \cdots, Y_n) = \sup_r \left| \frac{1}{r} \sum_{i=1}^{r} Y_i - \frac{1}{n-r} \sum_{i=r+1}^{n} Y_i \right|$$

$$= \sup_r \left| \frac{1}{r} \sum_{i=1}^{r} (a X_i + b) - \frac{1}{n-r} \sum_{i=r+1}^{n} (a X_i + b) \right|$$

$$= \sup_r \left| \frac{a}{r} \sum_{i=1}^{r} X_i + \frac{1}{r} \cdot rb - \frac{a}{n-r} \sum_{i=r+1}^{n} X_i - \frac{1}{n-r} (n-r)b \right|$$

$$= |a| \sup_r \left| \frac{1}{r} \sum_{i=1}^{r} X_i - \frac{1}{n-r} \sum_{i=r+1}^{n} X_i \right|$$

$$= |a| \quad S_n(X_1, \cdots, X_n) \ .$$

For given p and $\propto$, reject the null hypothesis for $Y_1, \cdots, Y_n$ that $\rho = n$ when $S_n(Y_1, \cdots, Y_n) \geq |a| \, s_\propto$ if and only if $S_n(X_1, \cdots, X_n) \geq s_\propto$ where $\propto \geq P\left[S_n(X_1, \cdots, X_n) \geq s_\propto | \rho\right]$.

For the more general problem where $Y_i \sim F(y; \theta)$ for $i = 1, \cdots, \rho$, $Y_i \sim F(y; \theta')$ for $i = \rho+1, \cdots, n$, and θ is a location parameter we can use our setup to construct a nonparametric test for detecting a change in the model. For example, suppose $Y_i \sim N(\mu_1, \sigma^2)$ for $i = 1, \cdots, \rho$, $Y_i \sim N(\mu_2, \sigma^2)$ for $i = p+1, \cdots, n$ and $\mu_1 > \mu_2$ where μ_1 and σ^2 are known and μ_2 may be unknown. To test $H_0: \rho = n, \mu_1$ vs. $H_1^+: \rho < n, \mu_1, \mu_2, \mu_1 > \mu_2$, transform Y_i by the rule

$$X_i = \begin{cases} 1 & \text{if } Y_i \geq \mu_1 \\ -1 & \text{if } Y_i < \mu_1 \end{cases} \ .$$

Using the statistic S_n^+, test

$$H_0': \rho = n, p \ \text{ vs. } \ H_1^{+'}: \rho < n, \ p, \ q, \ p > q$$

with $p = .5$ and $q = \int_{\mu_1}^{\infty} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(y-\mu_2)^2/2\sigma^2} dy$. If H_0' is

rejected, then reject H_0. If H_0 is rejected, estimate ρ as above and estimate μ_2 by $\frac{1}{n-\hat{\rho}} \sum_{i=\hat{\rho}+1}^{n} Y_i$.

1.4 Literature Review

The one-sided and two-sided tests for detecting a change point proposed above are intuitively appealing, but as one might suspect the properties of these tests are extremely tedious to evaluate. Unfortunately, it seems as though any test statistic one might consider as a reasonable candidate for testing for a change point suffers from the same drawback.

Consider for example the generalized likelihood ratio test. The likelihood function for $X_1, \cdots, X_n$ is

$$\prod_{i=1}^{n} p^{\frac{1+x_i}{2}} (1-p)^{\frac{1-x_i}{2}} \qquad \text{if } \rho = n,$$

$$\prod_{i=1}^{\rho} p^{\frac{1+x_i}{2}} (1-p)^{\frac{1-x_i}{2}} \prod_{i=\rho+1}^{n} q^{\frac{1+x_i}{2}} (1-q)^{\frac{1-x_i}{2}} \qquad \text{if } \rho \epsilon \{1, \cdots, n-1\}$$

and $$\prod_{i=1}^{n} q^{\frac{1+x_i}{2}} (1-q)^{\frac{1-x_i}{2}} \qquad \text{if } \rho = 0;$$

the generalized likelihood ratio when p and q are known is

$$\lambda = \frac{\prod_{i=1}^{n} p^{\frac{1+x_i}{2}} (1-p)^{\frac{1-x_i}{2}}}{\sup_r \left\{ \prod_{i=1}^{r} p^{\frac{1+x_i}{2}} (1-p)^{\frac{1-x_i}{2}} \prod_{i=r+1}^{n} q^{\frac{1+x_i}{2}} (1-q)^{\frac{1-x_i}{2}} \right\}}$$

$$= \inf_r \left\{ \prod_{i=r+1}^{n} \left(\frac{p}{q}\right)^{\frac{1+x_i}{2}} \left(\frac{1-p}{1-q}\right)^{\frac{1-x_i}{2}} \right\}$$

and $$c_{p,q} = c_{p,q}(x_1, \cdots, x_n) = 2 \log \lambda$$

1.4.1 $$= \inf_r \left\{ (n-r) \log \left(\frac{p}{q} \cdot \frac{1-p}{1-q}\right) + \left(\sum_{i=r+1}^{n} x_i\right) \log \left(\frac{p}{q} \cdot \frac{1-q}{1-p}\right) \right\} .$$

(When $r = n$, set the term in the brackets in the last equation equal to 0.) Both p and q are assumed known when performing the generalized likelihood ratio test H_0: $\rho = n, p$ *vs.* H_1: $\rho < n, p, q$. Evaluation of critical regions for this test statistic is considerably more complicated than for the test statistics we have proposed; note that $\sum_{i=r+1}^{n} X_i \in \{n-r, n-r-2, \cdots, -(n-r-2), -(n-r)\}$, $\log \left(\frac{p}{q} \cdot \frac{1-q}{1-p}\right) > 0$ if and only if $p > q$ and $\log \left(\frac{p}{q} \cdot \frac{1-p}{1-q}\right) > 0$ if and only if $\left|p - \frac{1}{2}\right| < \left|q - \frac{1}{2}\right|$.

Furthermore, tabulation of critical regions for the generalized likelihood ratio test for given value of n would take up a great deal of space even for a moderate number of values of p and q; the critical

regions for our test statistics depend only on p and hence can be written down more concisely.

Since the generalized likelihood ratio method often yields tests with good properties, it is conceivable that the effort required to evaluate critical regions may be worthwhile; we examine the test for $n = 5$. Since we are assuming both probabilities p and q are known, only the one-sided test is of interest. Furthermore, since $c_{p,q}(x_1, \cdots, x_n) = c_{1-p, 1-q}(-x_1, \cdots, -x_n)$, applying symmetry arguments to the one-sided test with $p<q$ will provide properties of the test for $p>q$. Thus, we will consider the likelihood ratio test for $p<q$ only. Table 1 lists the critical regions for the test with $p=.1$ and $q=.3, .9$ and $p=.5$ and $q=.7, .9$. Included in the listing for each p and q are the values c of the statistic in increasing order, the probability of the statistic being at least as small as c, the sequences yielding that value c, the point r in each sequence where c is attained, and the number of -1's in each sequence. (The point r where c is attained is usually uniquely defined; exceptions may occur when $p = 1-q$.) Note the following properties of the test evident from the equation for $c_{p,q}$ and suggested by the behavior of the test from Table 1.

For every $p<q$ and for all n, $c_{p,q} \leq 0$; $c_{p,q} = 0$ when $X_i = -1 \forall i$; $c_{p,q}$ attains its smallest value for the sequence where $x_i = 1 \forall i$ and the infinum occurs at $r = 0$. For any $p<q$ those

Table 1. The Generalized Likelihood Ratio Test Statistic, n=5.

c	$P(C_{.1,.3} \leq c)$	$x_1x_2x_3x_4x_5$	r	#-1's in $x_1, \cdots, x_5$
p = .1, q = .3				
-4.7715	.00001	11111	0	0
-3.8172	.00010	-11111	1	1
-3.5989	.00046	1111-1	0	1
		111-11	0	1
		11-111	0	1
		1-1111	0	1
-2.8629	.00127	-1-1111	2	2
-2.6446	.00370	-1111-1	1	2
		-111-11	1	2
		-11-111	1	2
-2.4263	.00856	111-1-1	0	2
		11-11-1	0	2
		1-111-1	0	2
		1-11-11	0	2
		11-1-11	0	2
		1-1-111	0	2
-1.9086	.01585	-1-1-111	3	3
-1.6903	.03043	-1-111-1	2	3
		-1-11-11	2	3
-1.472	.05230	-111-1-1	1	3
		-11-11-1	1	3
		-11-1-11	1	3
-1.2237	.08146	11-1-1-1	0	3
		1-11-1-1	0	3
		1-1-11-1	0	3
		1-1-1-11	0	3
-0.9543	.14707	-1-1-1-11	4	4
-0.7360	.21268	-1-1-11-1	3	4
-0.5172	.27829	-1-11-1-1	2	4
-0.2994	.34390	-11-1-1-1	1	4
-0.0811	.40951	1-1-1-1-1	0	4
0	1.00000	-1-1-1-1-1	5	5

Table 1 (Continued).

c	$P(C_{.1,.9} \leq c)$	$x_1, \cdots, x_5$	r	# -1's in $x_1, \cdots, x_5$
$p = .1,\ q = .9$				
-9.5425	.00001	11111	0	0
-7.9400	.00010	-11111	1	1
-5.7255	.00127	1111-1	0	1
		111-11	0	1
		11-111	0	1
		1-1111	0	1
		-1-1111	2	2
-3.8170	.04096	-1111-1	1	2
		-111-11	1	2
		1-1-111	3	2
		-11-111	1,3	2
		-1-1-111	3	3
-1.9085	.13978	111-1-1	0	2
		11-11-1	0	2
		1-111-1	0,2	2
		11-1-11	0,4	2
		1-11-11	0,2,4	2
		1-1-1-11	4	3
		-11-1-11	4	3
		-1-111-1	2	3
		-1-11-11	2,4	3
		-1-1-1-11	4	4
0	1.00000	11-1-1-1	5	3
		1-11-1-1	5	3
		-111-1-1	1,5	3
		1-1-11-1	3,5	3
		-11-11-1	1,3,5	3
		1-1-1-1-1	5	4
		-11-1-1-1	5	4
		-1-11-1-1	5	4
		-1-1-11-1	3,5	4
		-1-1-1-1-1	5	5

Table 1 (Continued).

c	$P(C_{.5,.7} \leq c)$	$x_1, \cdots, x_5$	r	# -1's in $x_1, \cdots, x_5$
$p = .5,\ q = .7$				
-1.4615	.03125	11111	0	1
-1.1692	.06250	-11111	1	1
-0.8769	.12500	1-1111	2	1
		-1-1111	2	2
-0.7255	.21875	1111-1	0	1
		111-11	0	1
		11-111	0	1
-0.5846	.31250	1-1-111	3	2
		-11-111	3	2
		-1-1-111	3	3
-0.4332	.37500	-1111-1	1	2
		-111-11	1	2
-0.2923	.50000	11-1-11	4	2
		1-11-11	4	2
		1-1-1-11	4	3
		-11-1-11	4	3
-0.1409	.62500	1-111-1	2	2
		-1-111-1	2	3
		-1-11-11	2	3
		-1-1-1-11	2	4
0	1.00000	111-1-1	5	2
		11-11-1	5	2
		11-1-1-1	5	3
		1-11-1-1	5	3
		-111-1-1	5	3
		1-1-11-1	5	3
		-11-11-1	5	3
		1-1-1-1-1	5	4
		-11-1-1-1	5	4
		-1-11-1-1	5	4
		-1-1-11-1	5	4
		-1-1-1-1-1	5	5

Table 1 (Continued).

c	$P(C_{.5,.9} \leq c)$	$x_1, \cdots, x_5$	r	# -1's in $x_1, \cdots, x_5$
p = .5, q = .9				
-2.5525	.03125	11111	0	0
-2.0420	.06250	-11111	1	1
-1.5315	.12500	1-1111	2	1
		-1-1111	2	2
-1.0210	.25000	11-111	3	1
		1-1-111	3	2
		-11-111	3	2
		-1-1-111	3	3
-0.6441	.31250	1111-1	0	1
		111-11	0	1
-0.5105	.53125	11-1-11	4	2
		1-11-11	4	2
		-111-11	4	2
		1-1-1-11	4	3
		-11-1-11	4	3
		-1-11-11	4	3
		-1-1-1-11	4	4
-0.1336	.56250	-1111-1	1	2
0	1.00000	111-1-1	5	2
		11-11-1	5	2
		1-111-1	5	2
		11-1-1-1	5	3
		1-11-1-1	5	3
		-111-1-1	5	3
		1-1-11-1	5	3
		-11-11-1	5	3
		-1-111-1	5	3
		1-1-1-1-1	5	4
		-11-1-1-1	5	4
		-1-11-1-1	5	4
		-1-1-11-1	5	4
		-1-1-1-1-1	5	5

sequences with fewer -1's tend to give smaller values of $c_{p,q}$ and those sequences with more -1's tend to give larger values of $c_{p,q}$. Also, those sequences with fewer -1's tend to have $c_{p,q}$ occurring at values of r near 0 and those sequences with more -1's tend to have $c_{p,q}$ occurring at values of r near the end of the sequence. This is somewhat alarming since when H_0 is rejected, an obvious estimate of the change point is that point at which $c_{p,q}$ is attained. In this case, estimates of change points would generally occur at the beginning of the sequence and detection of changes late in the sequence would be highly unlikely or impossible for any reasonable size-α test for many values of p and q. A closer examination of this dilemma will be made in later chapters. A comparison of the power of the test based on $C_{p,q}$ and S_n^- will be made in Chapter 3.

The generalized likelihood ratio test statistic $C_{p,q}$ has not been studied for finite sample sizes in the literature nor have tables been made available for its use in applications.

Asymptotic results using likelihood ratio test statistics for a change point problem for independent Bernoulli random variables have been obtained by D. Hinkley and E. Hinkley (1970), but their setup of the problem is somewhat different than ours. Their two-sided test hypothesis is H_0: $\rho = \rho_0$ versus H_1: $\rho \neq \rho_0$ where $\rho_0 < n$ and their one-sided alternatives are H_1': $\rho < \rho_0$ and H_1'': $\rho > \rho_0$ where $\rho_0 < n$. They assume that a change actually took place and are testing whether

the change took place at ρ_0; on the other hand we are testing whether any change has taken place, that is, if $\rho = n$. An iterative solution for the asymptotic distribution of their likelihood ratio test statistic is obtained using random walk results. For practical purposes, the solution is of limited use for two reasons. Firstly, a great deal of computation is required for finding critical regions for the tests by using the iterative formula. Secondly, the sample size n and the difference $n - \rho_0$ must be quite large for the asymptotic formulas to be accurate; when p and q are unknown, n and $n - \rho_0$ must be even larger in order to first estimate p and q, and then use the formula.

Possibly the most well-known and widely used statistic for testing for a change point is one proposed by E. S. Page (1955). His test statistic for the one-sided test with $p<q$ is

1.4.2 $$M = \max_{0 \le r \le n} \left[S_r - \min_{0 \le i < r} S_i \right]$$

where $S_r = \sum_{i=1}^{r} (X_i - (2p-1))$, $r = 1, \cdots, n$ and $S_0 = 0$. (The mean of X_i under the null hypothesis is $2p-1$; for other distributional setups, the same statistic M is used but S_r is replaced by $\sum_{i=1}^{r} (X_i - \theta)$ where $\theta = E(X_i \mid H_0)$.) The null hypothesis for a size-α test is rejected if $M \ge h_\alpha$ where h_α is such that $P(M \ge h_\alpha \mid p) \le \alpha$. Visually this amounts to graphing S_r vs. r and rejecting H_0 wherever an S_r rises a distance h_α above the previously plotted

minimum value on the graph. Intuitively one expects the graph to rise when an increase in the mean occurs and to remain horizontal when there is no change. When M occurs at a unique point r_0, then the estimator $\hat{\rho}$ of ρ is taken to be the most recent value where $\min_{0 \le i < r_0} S_i$ occurs; when M does not occur at a unique point, Page has not indicated a unique choice for the estimator $\hat{\rho}$.

Page has computed some critical regions for the test for the particular case where the X_i are symmetrical Bernoulli random variables and has computed the power of the test for one value of n. He has made no study of the types of sequences in the critical region for certain sizes of tests nor has he studied the properties of the estimator $\hat{\rho}$ of ρ. We consider the latter two properties crucial in evaluating change point tests. The case where the X_i are asymmetrical is also of interest.

To study these properties let us examine the test for $n = 5$. Evaluation of critical regions requires knowledge of p. These critical regions are more tedious to evaluate than for the test S_n^-. For $n = 5$ and $p = .1, .5, .9$, Table 2 lists the values m of the statistic, the probability that M is at least as large as m, the sequences yielding that value m, the point r in the sequence where m is attained, and the number of -1's in each sequence. Note the following properties of the test evident from the form of the test statistic and suggested by the behavior of the test from Table 2.

Table 2. Page's Statistic, n=5.

m	$P(M \geq m)$	$x_1, \cdots, x_5$	r	# -1's in $x_1, \cdots, x_5$
p = .1				
9.0	.00001	11111	0	0
7.2	.00019	-11111	1	1
		1111-1	0	1
7.0	.00046	1-1111	0	1
		11-111	0	1
		111-11	0	1
5.4	.00289	-1-1111	2	2
		-1111-1	1	2
		111-1-1	0	2
5.2	.00613	11-11-1	0	2
		1-111-1	0	2
		-111-11	1	2
		-11-111	1	2
5.0	.00856	11-1-11	0	2
		1-11-11	0	2
		1-1-111	0	2
3.6	.03772	11-1-1-1	0	3
		-111-1-1	1	3
		-1-111-1	2	3
		-1-1-111	3	3
3.4	.05959	1-1-11-1	0	3
		-11-1-11	1	3
		-1-11-11	2	3
3.2	.07417	1-1-11-1	0	3
		-11-1-11	1	3
3.0	.08146	1-1-1-11	0	3
1.8	.40951	1-1-1-1-1	0	4
		-11-1-1-1	1	4
		-1-11-1-1	2	4
		-1-1-11-1	3	4
		-1-1-1-11	4	4
0.0	1.00000	-1-1-1-1-1	0	5

Table 2 (Continued).

m	$P(M \geq m)$	$x_1, \cdots, x_5$	r	# -1's in $x_1, \cdots, x_5$
p = .3				
7.0	.00243	11111	0	0
5.6	.01377	-11111	1	1
		1111-1	0	1
5.0	.03078	1-1111	0	1
		11-111	0	1
		111-11	0	1
4.2	.07047	111-1-1	0	2
		-1111-1	1	2
		-1-1111	2	2
3.6	.12339	11-11-1	0	2
		1-111-1	0	2
		-111-11	1	2
		-11-111	1	2
3.0	.16308	11-1-11	0	2
		1-11-11	0	2
		1-1-111	0	2
2.8	.28656	-1-1-111	3	3
		-1-111-1	2	3
		-111-1-1	1	3
		11-1-1-1	0	3
2.2	.37917	1-11-1-1	0	3
		-11-11-1	1	3
		-1-11-11	2	3
1.6	.44091	1-1-11-1	0	3
		-11-1-11	1	3
1.4	.83193	1-1-1-11	0,4	3
		1-1-1-1-1	0	4
		-11-1-1-1	1	4
		-1-11-1-1	2	4
		-1-1-11-1	3	4
		-1-1-1-11	4	4
0.0	1.00000	-1-1-1-1-1	0	5

Table 2 (Continued).

m	P(M≥m)	$x_1, \cdots, x_5$	r	# -1's in $x_1, \cdots, x_5$
p = .5				
5.0	.03125	11111	0	0
4.0	.09375	1111-1	0	1
		-11111	1	1
3.0	.25000	111-11	0	1
		11-111	0	1
		1-1111	2	1
		111-1-1	0	2
		-1-1111	2	2
2.0	.59375	11-11-1	0	2
		1-111-1	2	2
		-1111-1	1	2
		11-1-11	0	2
		-111-11	1	2
		1-1-111	3	2
		-11-111	3	2
		11-1-1-1	0	3
		-111-1-1	1	3
		-1-111-1	2	3
		-1-1-111	3	3
1.0	.96875	1-11-11	0,2,4	2
		1-11-1-1	0,2	3
		1-1-11-1	0,3	3
		-11-11-1	1,3	3
		1-1-1-11	0,4	3
		-11-1-11	1,4	3
		-1-11-11	2,4	3
		1-1-1-1-1	0	4
		-11-1-1-1	1	4
		-1-11-1-1	2	4
		-1-1-11-1	3	4
		-1-1-1-11	4	4
0.0	1.00000	-1-1-1-1-1	0	5

For any p and n, $0 \leq \max_{0 \leq r \leq n} \left[S_r - \min_{0 \leq i < r} S_i \right] \leq 2n(1-p)$; the sequence of n -1's has $M = 0$ and the sequence of n 1's has $M = 2n(1-p)$. For any given value of p, those sequences with fewer -1's tend to give values of M near $2n(1-p)$ and those sequences with more -1's tend to give values of M near 0. In addition, for any given value of p, those sequences with fewer -1's tend to have M attained at values of r near 0 and those sequences with more -1's tend to have M attained at values of r near n. The unfortunate consequence is that estimates of change points will generally occur at the beginning of the sequence, making detection of later change points difficult. This same predicament was noted in connection with the generalized likelihood ratio statistic. The distribution of $\hat{\rho}$ for Page's statistic for $n = 5$ will be studied in later chapters. The power of this test will be calculated in Chapter 3.

A disturbing property of Page's test is that $X_1, \cdots, X_n$ and $X_n, \cdots, X_1$ give the same value of M. To see this, let $S_r(X_{i_1}, \cdots, X_{i_n}) = \sum_{i=i_1, \cdots, i_r} (X_i - (2p-1))$; for any $i < r$ we have $S_r(X_1, \cdots, X_n) - S_i(X_1, \cdots, X_n) = \{S_n(X_1, \cdots, X_n) - S_i(X_1, \cdots, X_n)\} - \{S_n(X_1, \cdots, X_n) - S_r(X_1, \cdots, X_n)\} = S_{n-i}(X_n, \cdots, X_1) - S_{n-r}(X_n, \cdots, X_1)$; the result now follows trivially. It is appropriate for a two-sided test for a change point to behave in such a manner,

but for the one-sided test one is interested in testing for a change in only one direction. It seems as though Page's statistic is actually testing something other than what is proposed.

1.5 Summary

Nonparametric one- and two-sided tests are proposed for detecting a change in the location parameter for a sequence of independent random variables. The aim is to obtain a test for small n which is easy to use and which will detect a change quickly.

As is characteristic of other change point tests, properties of the proposed tests are extremely tedious to evaluate. The method employed for deriving the distributions of the test statistics and change point random variables under the null and alternative hypotheses is to first obtain expressions for their joint distributions. Efficient procedures facilitating this task are presented. Unfortunately attempts for finding recursive formulas to study these distributions have been futile; large sample derivations are presented for the $\left[\frac{n}{2}\right] + 1$ largest values of the statistics, but for large n they are useful only when the size of the test is small.

Properties of the distributions of the test statistics under the null hypothesis are studied and are found to depend heavily on p ; critical regions for the one- and two-sided testing situations are tabulated. Techniques for testing for a change point when $n > 20$ are discussed.

Analyses of the change point random variables under the null hypothesis reveal the advantage of the proposed test statistics: whenever the null hypothesis is rejected each $\rho \in \{1, \cdots, n-1\}$ has positive probability of being estimated. Hence change points can be detected quickly. The generalized likelihood ratio test (1.4.1) and Page's test (1.4.2) both lack in this aspect: for each of these tests detection of changes later in the sequence is generally difficult or impossible.

CHAPTER 2

NULL HYPOTHESIS DISTRIBUTIONS

The distributions of the test statistics (1.2.2) and the corresponding change point random variables when the null hypothesis is true must be derived to find critical regions for the tests and to study properties of the estimator of the change point ρ. Obtaining any of these distributions is nontrivial, as one might suspect: the sample space of sequences $X_1, \cdots, X_n$ has 2^n elements! Fortunately, one need not generate all these sequences and then evaluate the statistics for each sequence to find the distributions; a more efficient procedure will be presented. The strategy will be to first obtain expressions for the joint distributions (S_n, R_n), (S_n^+, R_n^+), and (S_n^-, R_n^-), from which the marginal distributions are then derived and studied.

In the first section relevant sample spaces are defined and tools developed for employing the procedure; the counting procedure for finding the distributions is then presented. The marginal distributions S_n^+ and S_n are studied in Sections 3 and 4. In the next two sections the covariance between the test statistics and change point random variables is analyzed for the one- and two-sided testing situations. The change point random variables R_n and R_n^+ are investigated in Sections 7 and 8. Large sample derivations are presented in Section 9. The chapter is concluded with an example.

2.1 Definitions, Notations and Tools

Define the following sample spaces.

$$\Omega_n = \{x_1, \cdots, x_n : x_i = \pm 1 \text{ for } i = 1, \cdots, n\},$$

$$\Omega_{S_n, R_n} = \{(s, j): S_n(x_1, \cdots, x_n) = s, R_n(x_1, \cdots, x_n) = j$$

$$\text{where } x_1, \cdots, x_n \in \Omega_n\},$$

$$\Omega_{S_n} = \{s: S_n(x_1, \cdots, x_n) = s \text{ where } x_1, \cdots, x_n \in \Omega_n\},$$

and $\Omega_{R_n} = \{j: R_n(x_1, \cdots, x_n) = j \text{ where } x_1, \cdots, x_n \in \Omega_n\}$.

Similarly, define sample spaces for S_n^+, R_n^+, S_n^- and R_n^-.

$o(\Omega_n) = 2^n$ and for $x_1, \cdots, x_n \in \Omega_n$, $P(x_1, \cdots, x_n | p) = p^k(1-p)^{n-k}$ if k of the x_i's are 1 and $n-k$ of the x_i's are -1. There are $\binom{n}{k}$ sequences $x_1, \cdots, x_n$ with k 1's and $n-k$ -1's in Ω_n, $k = 0, \cdots, n$.

We have seen that s, $s^+ \in [0, 2]$ and $s^- \in [-2, 0]$, but unfortunately it appears that no simple scheme exists for listing the values for arbitrary n nor does a formula exist for $o(\Omega_{S_n})$, $o(\Omega_{S_n^+})$ or $o(\Omega_{S_n^-})$. Each of the latter quantities is an increasing function of n for $n \leq 20$ and $o(\Omega_{S_n}) \leq o(\Omega_{S_n^+}) = o(\Omega_{S_n^-})$. Even when the values attained by the statistics are known, no expression exists for their probabilities for given n. Explicit expressions for the most extreme values and then probabilities have been derived, but are of limited use unless the size α of the test is very small. Attempts for finding

recursive formulas to list more of the larger values of the test statistics and their probabilities for arbitrary n have been futile. The difficulty arises in being unable to obtain an ordered list of the values of the statistic; as n increases the manner in which the values are attained also increases and hence an ordering of the values as functions in n is impossible. The procedure presented in the next section will more clearly illustrate the nature of this difficulty.

For each $j \in \{0, 1, \cdots, n\}$ there is at least one sequence $x_1, \cdots, x_n \in \Omega_n$ such that $J_n(x_1, \cdots, x_n) = j$; thus $\Omega_{R_n} = \{0, 1, \cdots, n\}$. The same statement is true for $\Omega_{R_n^+}$ and $\Omega_{R_n^-}$.

Note that $\Omega_{S_n} \times \Omega_{R_n} \neq \Omega_{S_n, R_n}$. In general, for those sequences $x_1, \cdots, x_n$ with $S_n(x_1, \cdots, x_n) = s$, $R_n(x_1, \cdots, x_n)$ can assume only a few values of j. Similar statements hold for S_n^+, R_n^+, S_n^-, R_n^-.

The following notation will be needed. Define $c_k(s, j)$ to be the number of sequences $x_1, \cdots, x_n$ of k 1's and $n-k$ -1's that have $S_n(x_1, \cdots, x_n) = s$ and $R_n(x_1, \cdots, x_n) = j$. (If $x_1, \cdots, x_n$ has $J_n(x_1, \cdots, x_n) = j_1, \cdots, j_\ell$ then $x_1, \cdots, x_n$ contributes $\frac{1}{\ell}$ to $c_k(s, j_i)$, $i = 1, \cdots, \ell$.) Thus

2.1.1 $$P\{(S_n, R_n) = (s, j) \mid p\} = \sum_{k=0}^{n} c_k(s, j)\, p^k (1-p)^{n-k} .$$

Define $c_k(s) = \sum_{j=0}^{n} c_k(s, j)$ so that

2.1.2 $$P(S_n = s \mid p) = \sum_{k=0}^{n} c_k(s)\, p^k (1-p)^{n-k}$$

and $C_k(s_{\alpha, p}) = \sum_{s \geq s_{\alpha, p}} c_k(s)$ so that $s_{\alpha, p}$ is the smallest value of s satisfying

2.1.3 $$\alpha \geq \sum_{k=0}^{n} C_k(s_{\alpha, p})\, p^k (1-p)^{n-k} \ .$$

Define $d_k(j) = \sum_{s \in \Omega_{S_n}} c_k(s, j)$ so that

2.1.4 $$P(R_n = j \mid p) = \sum_{k=0}^{n} d_k(j)\, p^k (1-p)^{n-k} \ .$$

Similarly, define $c_k^+(s^+, j)$, $c_k^+(s^+)$, $C_k^+(s_{\alpha, p}^+)$, $d_k^+(j)$, $c_k^-(s^-, j)$, $c_k^-(s^-)$, $C_k^-(s_{\alpha, p}^-)$ and $d_k^-(j)$.

The procedure for obtaining the joint distributions involves counting the $c_k(s, j)$'s and $c_k^+(s^+, j)$'s. We establish some results useful for that purpose. The first three results deal with the random variables associated with the one-sided test H_0 vs. H_1^+.

<u>Lemma 1.</u> For each $n \geq 1$ and $x_1, \cdots, x_n \in \Omega_n$, $S_n^+(x_1, \cdots, x_n) = S_n^+(-x_n, \cdots, -x_1)$; if $S_n^+(x_1, \cdots, x_n)$ occurs at $r = j$ then $S_n^+(-x_n, \cdots, -x_1)$ occurs at $r = n-j$.

Proof. Note that $\frac{1}{r}\sum_{i=1}^{r} x_i - \frac{1}{n-r}\sum_{i=r+1}^{n} x_i = \frac{1}{n-r}\sum_{i=n}^{r+1}(-x_i)$ $- \frac{1}{r}\sum_{i=r}^{1}(-x_i)\ \forall r$; hence $S_n^+(x_1, \cdots, x_n) = \sup_r \{\frac{1}{r}\sum_{i=1}^{r} x_i$ $- \frac{1}{n-r}\sum_{i=r+1}^{r} x_i\} = \sup_r \{\frac{1}{n-r}\sum_{i=n}^{r+1}(-x_i) - \frac{1}{r}\sum_{i=r}^{1}(-x_i)\} =$ $S_n^+(-x_n, \cdots, -x_1)$. *Also note* that if $S_n^+(x_1, \cdots, x_n)$ occurs at $r = j$ then $S_n^+(-x_n, \cdots, -x_1)$ occurs at $r = n-j$. ■

Lemma 1 provides the basis for a number of useful results.

Theorem 2. For each $n \in Z^+$,

(i) $c_k^+(s^+, j) = c_{n-k}^+(s^+, n-j)$, $c_k^+(s^+) = c_{n-k}^+(s^+)$ and $C_k^+(s_{\alpha, p}^+) = C_{n-k}^+(s_{\alpha, p}^+) = C_k^+(s_{\alpha, 1-p}^+)\ \forall s^+ \in \Omega_{S_n^+}$, $j \in \Omega_{R_n^+}$, $p \in [0, 1]$ and $k \in \{0, 1, \cdots, n\}$;

(ii) $P(S_n^+ = s^+ \mid p) = P(S_n^+ = s^+ \mid 1-p)\ \forall s^+ \in \Omega_{S_n^+}$ and $p \in [0, 1]$;

(iii) $d_k^+(j) = d_{n-k}^+(n-j) = d_k^-(j) = d_{n-k}^-(n-j)\ \forall\, j, k \in \{0, 1, \cdots, n\}$;

and (iv) $P(R_n^+ = j \mid p) = P(R_n^+ = n-j \mid 1-p)\ \forall\, j \in \Omega_{R_n^+}$ and $p \in [0, 1]$.

Proof. (i) Refer to the proof of Lemma 1 and note that if $x_1, \cdots, x_n$ has k 1's then $-x_n, \cdots, -x_1$ has $n-k$ 1's .

(ii) $P(S_n^+ = s^+ \mid p) = \sum_{k=0}^{n} c_k^+(s^+)\, p^k (1-p)^{n-k} =$

$\sum_{k=0}^{n} c_{n-k}^+(s^+)\,(1-p)^{n-k}\, p^k = P(S_n^+ = s^+ \mid 1-p)$.

(iii) Note that $d_k^+(j) = \sum_{s^+ \in \Omega_{S_n^+}} c_k^+(s^+, j)$ and $c_k^+(s^+, j) =$

$c_{n-k}^+(s^+, n-j)$.

(iv) $P(R_n^+ = j \mid p) = \sum_{k=0}^{n} d_k^+(j)\, p^k (1-p)^{n-k}$

$= \sum_{k=0}^{n} d_{n-k}^+(n-j)\, p^k (1-p)^{n-k} = \sum_{k=0}^{n} d_k^+(n-j)\,(1-p)^k\, p^{n-k}$

$= P(R_n^+ = n-j \mid 1-p)$. ■

Note that (i) implies that we need only find the $c_k^+(s^+, j)$'s *for* $k \le \left[\frac{n}{2}\right]$ or for $j \le \left[\frac{n}{2}\right]$ to evaluate the joint distribution. The numerical computation of critical regions need only be carried out for $p \in (0, \frac{1}{2}]$ by (ii). The distribution of R_n^+ is symmetrical in j for $p = \frac{1}{2}$ by (iv).

The next theorem provides upper and lower bounds for the values $s^+ \in \Omega_{S_n^+}$ as a function of the number of 1's in the sequence.

<u>Theorem 3.</u> For any $n \ge 2$ and $k \in \{1, \cdots, n-1\}$, max $S_n^+(X_1, \cdots, X_n) = 2$ and min $S_n^+(X_1, \cdots, X_n) = \left|\frac{n-2k}{n}\right|$ where $X_1, \cdots, X_n$ is any sequence with k 1's ; furthermore there exist sequences of k 1's which attain this maximum and minimum.

Proof. For $k \epsilon \{1, \cdots, n-1\}$ consider the sequence $x_1, \cdots, x_n$ with $x_i = 1$ for $i = 1, \cdots, k$ and $x_i = -1$ for $i = k+1, \cdots, n$; $S_n^+(x_1, \cdots, x_n) = 2$. For $k \epsilon \{1, \cdots, n-1\}$ consider the sequence $x_1, \cdots, x_n$ with $x_i = -1$ for $i = 1, \cdots, n-k$ and $x_i = 1$ for $i = n-k+1, \cdots, n$. If $k < \left[\frac{n}{2}\right]$ then $R_n^+(x_1, \cdots, x_n) = 0$ and $S_n^+(x_1, \cdots, x_n) = \frac{n-2k}{n}$; if $k > \left[\frac{n}{2}\right]$ then $R_n^+(x_1, \cdots, x_n) = n$ and $S_n^+(x_1, \cdots, x_n) = \frac{2k-n}{n}$; if $k = \frac{n}{2}$ then $J_n^+(x_1, \cdots, x_n) = 0$ or n and $S_n^+(x_1, \cdots, x_n) = 0$ ■

Theorem 4. Lemma 1 is true with S_n^- replacing S_n^+. Theorem 2 is true when the '+' superscripts are replaced by '-' superscripts. Theorem 3 is true with $\min S_n^-(X_1, \cdots, X_n) = -2$ and $\max S_n^-(X_1, \cdots, X_n) = -\left|\frac{n-2k}{n}\right|$.

Proof. Note that $S_n^+(x_1, \cdots, x_n) = \sup_r \{\frac{1}{r} \sum_{i=1}^{r} x_i - \frac{1}{n-r} \sum_{i=r+1}^{n} x_i\}$

$$= -\inf_r \{\frac{1}{r} \sum_{i=1}^{r} (-x_i) - \frac{1}{n-r} \sum_{i=r+1}^{n} (-x_i)\} = S_n^-(-x_1, \cdots, -x_n)$$

$\forall x_1, \cdots, x_n \epsilon \Omega_n$. ■

Note that $c_k^+(s^+, j) = c_{n-k}^-(-s^+, j)$, $c_k^+(s^+) = c_{n-k}^-(-s^+)$, $C_k^+(s^+_{,p}) = C_{n-k}^+(-s^+_{,p})$ and $P(S_n^+ = s^+ \mid p) = P(S_n^- = -s^+ \mid p)$ $\forall s^+ \epsilon \Omega_{S_n^+}$, $j \epsilon \Omega_{R_n^+}$, $k \epsilon \{0_1, \cdots, n\}$ $p \epsilon [0, 1]$ and $n \epsilon Z^+$.

Due to the perfect symmetry between the tests H_0 vs. H_1^+ and H_0 vs. H_1^- and the distributions related to these tests, any statement

made about the former can be correspondingly made about the latter. Hence, we will discuss only the former in what follows.

Results similar to Lemma 1 and Theorem 2 are presented next for the random variables associated with the two-sided test.

<u>Lemma 5.</u> For each $n \geq 1$ and $x_1, \cdots, x_n \in \Omega_n$, $S_n(x_1, \cdots, x_n) = S_n(-x_1, \cdots, -x_n) = S_n(x_n, \cdots, x_1) = S_n(-x_n, \cdots, -x_1)$; if $S_n(x_1, \cdots, x_n)$ occurs at $r = j$ then $S_n(-x_1, \cdots, -x_n)$ occurs at $r = j$, and $S_n(x_n, \cdots, x_1)$ and $S_n(-x_n, \cdots, -x_1)$ occur at $r = n-j$.

<u>Proof.</u> $\left| \frac{1}{r} \sum_{i=1}^{r} x_i - \frac{1}{n-r} \sum_{i=r+1}^{n} x_i \right| = \left| \frac{1}{r} \sum_{i=1}^{r} (-x_i) - \frac{1}{n-r} \sum_{i=r+1}^{n} (-x_i) \right|$ $\forall r$ and $\left| \frac{1}{r} \sum_{i=1}^{r} x_i - \frac{1}{n-r} \sum_{i=r+1}^{n} x_i \right| = \left| \frac{1}{n-r} \sum_{i=n}^{r+1} x_i - \frac{1}{r} \sum_{i=r}^{1} x_i \right| \; \forall r.$ ∎

<u>Theorem 6.</u> For each $n \in Z^+$,

(i) $c_k(s, j) = c_k(s, n-j) = c_{n-k}(s, j) = c_{n-k}(s, n-j)$, $c_k(s) = c_{n-k}(s)$ and $C_k(s_\alpha, p) = C_{n-k}(s_\alpha, p) = C_k(s_\alpha, 1-p)$ $\forall s \in \Omega_{S_n}$, $j \in \Omega_{R_n}$, $p \in [0, 1]$ and $k \in \{0, 1, \cdots, n\}$;

(ii) $P(S_n = s \mid p) = P(S_n = s \mid 1-p) \; \forall s \in \Omega_{S_n}$ and $p \in [0, 1]$;

(iii) $d_k(j) = d_{n-k}(j) = d_{n-k}(n-j) = d_k(n-j) \; \forall j$, $k \in \{0, 1, \cdots, n\}$;

and (iv) $P(R_n = j \mid p) = P(R_n = n-j \mid p) = P(R_n = j \mid 1-p) = P(R_n = n-j \mid 1-p) \; \forall j \in \Omega_{R_n}$ and $p \in [0, 1]$.

The proof of Theorem 6 follows directly from Lemma 5 and the definitions.

By (i) all of the $c_k(s,j)$'s can be found by evaluating those for which $k \leq \left[\frac{n}{2}\right]$ and $j \leq \left[\frac{n}{2}\right]$. By (ii) numerical computation of critical regions need be carried out only for $p \in (0, \frac{1}{2}]$. The distribution of R_n is symmetrical in j and p by (iv).

The upper and lower bounds for values in Ω_{S_n} are as follows.

<u>Theorem 7.</u> For $n \geq 2$ and $k \in \{1, \cdots, n-1\}$, max $S_n(X_1, \cdots, X_n) = 2$ where $X_1, \cdots, X_n$ is any sequence with k -1's; for each k there exists a sequence which attains this maximum.

For $k \in \{1, \cdots, n-1\}$ let $L_e(k)$ and $L_o(k)$ denote min $S_n(X_1, \cdots, X_n)$ when n is even or odd, respectively, and $X_1, \cdots, X_n$ is any sequence with k -1's. We have

$$L_e(k) = L_e(n-k) = \frac{n-2k}{n} \quad \text{for } k = 1, \cdots, \left[\frac{n-1}{4}\right],$$

$$L_e(k) = L_e(n-k) = \frac{n}{2(n-2)} \quad \text{for } k = \frac{n}{4},$$

$$L_e(k) = L_e(n-k) = \frac{2k}{n-1} \quad \text{for } k = \left[\frac{n}{4}\right] + 1, \cdots, \frac{n}{2},$$

$$L_o(k) = L_o(n-k) = \frac{n-2k}{n} \quad \text{for } k = 1, \cdots, \left[\frac{n-2}{4}\right],$$

$$L_o(k) = L_o(n-k) = \frac{n-1}{2(n-2)} \quad \text{for } k = \frac{n-1}{4},$$

and $$L_o(k) = L_o(n-k) = \frac{2k}{n-1} \quad \text{for } k = \left[\frac{n+3}{4}\right], \cdots, \frac{n-1}{2}.$$

For each k the lower bound is attainable.

Proof. We establish the lower bounds for n even only; the method of proof for n odd is similar.

$L_e(k) = L_e(n-k)$ for $k \in \{0, 1, \cdots, n\}$ by Lemma 5.

Note that if $k \le \left[\frac{n}{2}\right]$ and $x_1 = -1$ then $S_n(-1, x_2, \cdots, x_n) \ge 1$.

We are interested in finding lower bounds so in what follows assume $x_1 = 1$.

Suppose $k < \left[\frac{n-1}{4}\right]$; $\left| \frac{1}{n} \sum_{i=1}^{n} x_i \right| = \frac{n-2k}{n}$ so $L_e(k) \ge \frac{n-2k}{n}$.

Suppose $k = \frac{n}{4}$. If $x_2 = 1$ then $\left| \frac{1}{2} \sum_{i=1}^{2} x_i - \frac{1}{n-2} \sum_{i=3}^{n} x_i \right| = \left| 1 - \frac{n-2-2k}{n-2} \right| = \frac{2k}{n-2} = \frac{n}{2(n-2)}$. If $x_2 = 1$ then $\left| \frac{1}{2} \sum_{i=1}^{2} x_i - \frac{1}{n-2} \sum_{i=3}^{n} x_i \right| = \left| 0 - \frac{n-2-2(k-1)}{n-2} \right| = \frac{n-2k}{n-2} = \frac{n}{2(n-2)}$.

Hence $L_e\left(\frac{n}{4}\right) \ge \frac{n}{2(n-2)}$.

Suppose $\left[\frac{n}{4}\right] + 1 \le k \le \frac{n}{2}$; $\left| \frac{1}{1} \sum_{i=1}^{1} x_i - \frac{1}{n-1} \sum_{i=2}^{n} x_i \right| = \left| 1 - \frac{n-1-2k}{n-1} \right| = \frac{2k}{n-1}$. Hence $L_e(k) \ge \frac{2k}{n-1}$.

We show that for $k < \left[\frac{n-1}{4}\right]$ that $L_e(k)$ is attainable. If k is even consider the sequence with $x_i = 1$ for $i = 1, \cdots, \frac{n-k}{2}$, $x_i = -1$ for $i = \frac{n-k}{2} + 1, \cdots, \frac{n+k}{2}$ and $x_i = 1$ for $i = \frac{n+k}{2} + 1, \cdots, n$. For this sequence S_n may occur only for $r = 0, \frac{n-k}{2}, \frac{n+k}{2}, n$ by Theorem 15. To verify that R_n for this sequence occurs at $r = 0$ is straightforward. If k is odd construct a sequence with $x_i = 1$ for

$i = 1, \cdots, \frac{n-k-1}{2}$, $x_i = -1$ for $i = \frac{n-k-1}{2} + 1, \cdots, \frac{n+k-1}{2}$ and $x_i = 1$ for $i = \frac{n+k+1}{2}, \cdots, n$.

Lastly, we show that for $k = \frac{n}{4}, \left[\frac{n}{4}\right] + 1, \cdots, \frac{n}{2}$ $L_e(k)$ is attainable. Construct the following sequence: $x_1 = x_n = 1$, $x_2 = x_{n-1} = -1$, $x_i = 1$ for $i = 3, \cdots, \frac{n}{2} - \left[\frac{k-1}{2}\right]$ and $i = \frac{n}{2} + \left[\frac{k-1}{2}\right] + 1$, $\cdots$, n-2 , and if k is odd $x_i = 1$ for $i = \frac{n}{2}$ and $x_i = 1$ for all remaining i or if k is even $x_i = -1$ for $i = \frac{n}{2} - \left[\frac{k-1}{2}\right] + 1, \cdots, \frac{n}{2} + \left[\frac{k-1}{2}\right]$. S_n for this sequence may occur only for $r = 0, n, 1, n-1, 2, n-2, \frac{n}{2} - \left[\frac{k-1}{2}\right], \frac{n}{2} + \left[\frac{k-1}{2}\right]$ if k is even and also for $r = \frac{n}{2} - 1, \frac{n}{2}$ if k is odd . ■

To obtain Ω_{S_n} and $\Omega_{S_n^+}$ our intention is to construct A_n as defined in the next theorem.

Theorem 8. $\Omega_{S_n} \subset \Omega_{S_n^+} = A_n$ *for each* $n \geq 1$ where

$$A_n = \{ \frac{1}{r} \sum_{i=1}^{r} x_i - \frac{1}{n-r} \sum_{i=r+1}^{n} x_i : x_1, \cdots, x_n \in \Omega_n \text{ and if }$$

$$x_1, \cdots, x_n \text{ has } k \text{ 1's then } \frac{1}{r} \sum_{i=1}^{r} x_i - \frac{1}{n-r} \sum_{i=r+1}^{n} x_i \geq \left| \frac{n-2k}{n} \right| ,$$

$$k = 0, \cdots, n \text{ and } r = 0, \cdots, \left[\frac{n}{2}\right] \} .$$

Proof. For any sequence $x_1, \cdots, x_n \in \Omega_n$

$$S_n(x_1, \cdots, x_n) = \sup_r \left| \frac{1}{r} \sum_{i=1}^{r} x_i - \frac{1}{n-r} \sum_{i=r+1}^{n} x_i \right|$$

$$= \max \{ \sup_r \left[\frac{1}{r} \sum_{i=1}^{r} x_i - \frac{1}{n-r} \sum_{i=r+1}^{n} x_i \right],$$

$$\sup_r \left[\frac{1}{r} \sum_{i=1}^{r} (-x_i) - \frac{1}{n-r} \sum_{i=r+1}^{n} (-x_i) \right] \}$$

$$= \max \{ S_n^+(x_1, \cdots, x_n), \; S_n^+(-x_1, \cdots, -x_n) \} .$$

Thus $\Omega_{S_n} \subset \Omega_{S_n^+}$.

If $S_n^+(x_1, \cdots, x_n)$ occurs at some $j > \left[\frac{n}{2}\right]$ then $S_n^+(-x_n, \cdots, -x_1)$ occurs at $n-j \leq \left[\frac{n}{2}\right]$ and $S_n^+(-x_n, \cdots, -x_1) \in A_n$, but $S_n^+(x_1, \cdots, x_n) = S_n^+(-x_n, \cdots, -x_1)$. Hence $\Omega_{S_n^+} \subseteq A_n$.

Next, we show that $A_n \subseteq \Omega_{S_n^+}$. We must show that for each value $s_0 = \frac{1}{r} \sum_{i=1}^{r} x_i - \frac{1}{n-r} \sum_{i=r+1}^{n} x_i \geq \left| \frac{n-2k}{n} \right|$ where $x_1, \cdots, x_n$ has

k 1's that there exists a sequence with k 1's satisfying $S_n^+ = s_0$. If $x_1, \cdots, x_n$ has a 1's in the first r places and n-k 1's in the last n-r places, consider the sequence $y_1, \cdots, y_n$ with $y_i = -1$ for $i = 1, \cdots, r-a$, $y_i = 1$ for $i = r-a+1, \cdots, r$, $y_i = -1$ for $i = r+1, \cdots, n-k$ and $y_i = 1$ for $i = n-k+1, \cdots, n$. $S_n^+(y_1, \cdots, y_n) = s_0$. ∎

The set inclusion $\Omega_{S_n} \subset \Omega_{S_n^+}$ is proper for $n = 2, \cdots, 20$. If $s^+ \epsilon\, \Omega_{S_n^+} \setminus \Omega_{S_n}$ and if s^+ occurs for a sequence with k 1's then $s^+ < s$ for all s which occur for a sequence of k 1's; $s^+ \epsilon\, \Omega_{S_n^+} \setminus \Omega_{S_n}$ does not imply $s^+ < s \;\forall s \,\epsilon\, \Omega_{S_n}$.

One might suspect that for $s \,\epsilon\, \Omega_{S_n}$, $\frac{1}{2}\, c_k(s) \doteq c_k^+(s)$. This relationship holds for some k and s, however, the set of such values is not large enough to be useful.

2.2 The Procedure

The procedure used to find the joint and marginal distributions is now presented.

Notation. We need to be able to describe types of sequences $x_1, \cdots, x_n$. By the expression $T_n^k(r, a)$ denote the set of all those sequences of length n with k 1's where a 1's occur in the *first* r places and k-a 1's occur in the last n-r places.

For example $T_5^3(2,1) = \{1\text{-}111\text{-}1,\ 1\text{-}11\text{-}11,\ 1\text{-}1\text{-}111,\ \text{-}1111\text{-}1,$ $\text{-}111\text{-}11,\ \text{-}11\text{-}111\}$.

The Procedure

I. List the values $\frac{1}{r} \sum_{i=1}^{r} x_i - \frac{1}{n-r} \sum_{i=r+1}^{n} x_i \in A_n$ for each r systematically by varying the number of -1's in the sequence. For each value include the types of sequences yielding that value.

For $r = 0$ we have the following list.

$0 - \frac{1}{n} \sum_{i=1}^{n} x_i$	Types of Sequences
1	$T_n^0(0,0)$
$\frac{n-2}{n}$	$T_n^1(0,0)$
$\frac{n-4}{n}$	$T_n^2(0,0)$
$\vdots$	
0	$T_n^{\frac{n}{2}}(0,0)$ if n is even
$\frac{1}{n}$	$T_n^{\frac{n-1}{2}}(0,0)$ if n is odd

For $r = 1$ we have the following list.

$\frac{1}{1} x_1 - \frac{1}{n-1} \sum_{i=2}^{n} x_i$	Types of Sequences
$1 - (-\frac{n-1}{n-1}) = 2$	$T_n^1 (1, 1)$
$1 - (-\frac{n-3}{n-1}) = \frac{2(n-2)}{n-1}$	$T_n^2 (1, 1)$
$1 - (-\frac{n-5}{n-1}) = \frac{2(n-3)}{n-1}$	$T_n^3 (1, 1)$
$\vdots$	
$\left\{ \begin{array}{l} 1 - (-\frac{1}{n-1}) = \frac{n}{n-1} \\ \quad 1 \end{array} \right.$	$T_n^{\frac{n}{2}} (1, 1)$ if n is even $T_n^{\frac{n+1}{2}} (1, 1)$ if n is odd
$\left\{ \begin{array}{l} 1 - \frac{1}{n-1} = \frac{n-2}{n-1} \\ 1 - \frac{2}{n-1} = \frac{n-3}{n-1} \end{array} \right.$	$T_n^{\frac{n+2}{2}} (1, 1)$ if n is even $T_n^{\frac{n+3}{2}} (1, 1)$ if n is odd
$\vdots$	

For $r = 2$ we have the following lists.

$\frac{1}{2}\sum_{i=1}^{2} x_i - \frac{1}{n-2}\sum_{i=3}^{n} x_i$	Types of Sequences
$1 - (-\frac{n-2}{n-2}) = 2$	$T_n^2\,(2,2)$
$1 - (-\frac{n-4}{n-2}) = \frac{2(n-3)}{n-2}$	$T_n^3\,(2,2)$
$1 - (-\frac{n-6}{n-2}) = \frac{2(n-4)}{n-2}$	$T_n^4\,(2,2)$
$\cdot$	
$\cdot$	
$\cdot$	
1	$T_n^{\frac{n+2}{2}}\,(2,2)$ if n is even
$1 - (-\frac{1}{n-2}) = \frac{n-1}{n-2}$	$T_n^{\frac{n+1}{2}}\,(2,2)$ if n is odd
$1 - \frac{2}{n-2} = \frac{n-4}{n-2}$	$T_n^{\frac{n+4}{2}}\,(2,2)$ if n is even
$1 - \frac{1}{n-2} = \frac{n-3}{n-2}$	$T_n^{\frac{n+3}{2}}\,(2,2)$ if n is odd
$\cdot$	
$\cdot$	
$\cdot$	

$\frac{1}{2}\sum_{i=1}^{2} x_i - \frac{1}{n-2}\sum_{i=3}^{n} x_i$	Types of Sequences
$0 - (-\frac{n-2}{n-2}) = 1$	$T_n^1(2,1)$
$\frac{n-4}{n-2}$	$T_n^2(2,1)$
$\frac{n-6}{n-2}$	$T_n^3(2,1)$
$\cdot$	
$\cdot$	
$\cdot$	
$\begin{cases} 0 \\ \frac{1}{n-2} \end{cases}$	$T_n^{\frac{n}{2}}(2,1)$ if n is even $T_n^{\frac{n-1}{2}}(2,1)$ if n is odd

For $r = 3$ we have the following lists.

$\frac{1}{3}\sum_{i=1}^{3} x_i - \frac{1}{n-3}\sum_{i=4}^{n} x_i$	Types of Sequences
$1 - (-\frac{n-3}{n-3}) = 2$	$T_n^3(3,3)$
$1 - (-\frac{n-5}{n-3}) = \frac{2(n-4)}{n-3}$	$T_n^4(3,3)$
$1 - (-\frac{n-7}{n-3}) = \frac{2(n-5)}{n-3}$	$T_n^5(3,3)$
$\cdot$	
$\cdot$	
$\cdot$	

$$\begin{cases} 1 - \left(-\dfrac{1}{n-3}\right) = \dfrac{n-2}{n-3} & T_n^{\frac{n+2}{2}}(3,3) \quad \text{if } n \text{ is even} \\ 1 & T_n^{\frac{n+3}{2}}(3,3) \quad \text{if } n \text{ is odd} \end{cases}$$

$$\begin{cases} 1 - \dfrac{1}{n-3} = \dfrac{n-4}{n-3} & T_n^{\frac{n+4}{2}}(3,3) \quad \text{if } n \text{ is even} \\ 1 - \dfrac{2}{n-3} = \dfrac{n-5}{n-3} & T_n^{\frac{n+5}{2}}(3,3) \quad \text{if } n \text{ is odd} \end{cases}$$

$\vdots$

$\frac{1}{3}\sum_{i=1}^{3} x_i - \frac{1}{n-3}\sum_{i=4}^{n} x_i$	*Types of Sequences*
$\frac{1}{3} - \left(-\frac{n-3}{n-3}\right)$	$T_n^2(3,2)$
$\frac{1}{3} - \left(-\frac{n-5}{n-3}\right)$	$T_n^3(3,2)$

$\vdots$

$$\begin{cases} \dfrac{1}{3} - \left(-\dfrac{1}{n-3}\right) & T_n^{\frac{n}{2}}(3,2) \quad \text{if } n \text{ is even} \\ \dfrac{1}{3} & T_n^{\frac{n+1}{2}}(3,2) \quad \text{if } n \text{ is odd} \end{cases}$$

$$\begin{cases} \dfrac{1}{3} - \dfrac{1}{n-3} & T_n^{\frac{n+2}{2}}(3,2) \quad \text{if } n \text{ is even} \\ \dfrac{1}{3} - \dfrac{2}{n-3} & T_n^{\frac{n+3}{2}}(3,2) \quad \text{if } n \text{ is odd} \end{cases}$$

$\vdots$

$\frac{1}{3}\sum_{i=1}^{3} x_i - \frac{1}{n-3}\sum_{i=4}^{n} x_i$	Types of Sequences
$-\frac{1}{3} - (-\frac{n-3}{n-3})$	$T_n^1(3, 1)$
$-\frac{1}{3} - (-\frac{n-5}{n-3})$	$T_n^2(3, 1)$
$-\frac{1}{3} - (-\frac{n-7}{n-3})$	$T_n^3(3, 1)$
$\vdots$	

The method of listing must be continued in the above manner for each $r \leq \left[\frac{n}{2}\right]$.

By inspection similar lists for $r > \left[\frac{n}{2}\right]$ can be found from those for $r \leq \left[\frac{n}{2}\right]$ and hence need not be written down. For example, for $r = 1$ the value $\frac{1}{1}\sum_{i=1}^{1} x_i - \frac{1}{n-1}\sum_{i=2}^{n} x_i = \frac{2(n-2)}{n-1}$ may occur for the types of sequences $T_n^2(1, 1)$; for $r = n-1$ the value $\frac{1}{n-1}\sum_{i=1}^{n-1} x_i - \frac{1}{1}\sum_{i=n}^{n} x_i = \frac{2(n-2)}{n-1}$ may occur for the types of sequences $T_n^{n-2}(n-1, n-2)$. See Lemma 1.

II. For given n, categorize the information found in I. according to the number of -1's in the sequence for # -1's = $0, \cdots, \left[\frac{n}{2}\right]$. That is, for # -1's = 0 list all values for the sequences having 0 -1's; include the corresponding point r and the types of sequences giving

that value. The information for $r > \left[\frac{n}{2}\right]$ as found from I. by inspection must be included. Do the same for # -1's = $1, 2, \cdots, \left[\frac{n}{2}\right]$.

<u>Example 1</u>. The procedure for $n = 5$ is demonstrated.

 # -1's = 0

s^+	r	Types of Sequences
1	5	$T_5^5\,(5, 5)$

 # -1's = 1

s^+	r	Types of Sequences
$\frac{3}{5}$	5	$T_5^4\,(5, 4)$
2	4	$T_5^4\,(4, 4)$
$\frac{2}{3}$	2	$T_5^4\,(2, 2)$
1	3	$T_5^4\,(3, 3)$

 # -1's = 2

s^+	r	Types of Sequences
$\frac{1}{5}$	5	$T_5^3\,(5, 3)$
1	1	$T_5^3\,(1, 1)$
$\frac{3}{2}$	4	$T_5^3\,(4, 3)$
$\frac{4}{3}$	2	$T_5^3\,(2, 2)$
2	3	$T_5^3\,(3, 3)$
$\frac{1}{3}$	3	$T_5^3\,(3, 2)$

Note that for # -1's = $\left[\frac{n}{2}\right] + 1, \cdots, n$ similar listings can be found from those for # -1's = $0, 1, \cdots, \left[\frac{n}{2}\right]$.

III. For given n the $c_k^+(s^+, j)$'s and $c_k(s, j)$'s are now counted; one need only count them for # -1's = 0, 1, $\cdots$, $\left[\frac{n}{2}\right]$ by Theorems 2 and 6 .

Order the values in II. within each category of # -1's from the largest to the smallest; within each category count the number of sequences yielding each value. Include the point j at which the value is attained and the types of sequences yielding that value.

<u>Example 2</u>. For n = 5 the $c_k^+(s^+, j)$'s are found for # -1's = 0, 1, 2.

-1's = 0

s^+	j	$c_5^+(s^+, j)$	Types of Sequences
1	5	1	$T_5^5(5, 5)$

-1's = 1

s^+	j	$c_4^+(s^+, j)$	Types of Sequences
2	4	1	$T_5^4(4, 4)$
1	3	1	$T_5^4(3, 3)$
$\frac{2}{3}$	2	1	$T_5^4(2, 2)$
$\frac{3}{5}$	5	2	$T_5^4(5, 4)$

-1'2 = 2

s^+	j	$c_3^+(s^+, j)$	Types of Sequences
2	3	1	$T_5^3(3, 3)$
$\frac{3}{2}$	4	3	$T_5^3(4, 3)$
$\frac{4}{3}$	2	1	$T_5^3(2, 2)$
1	1	2	$T_5^3(1, 1)$
$\frac{1}{3}$	3	1	$T_5^3(3, 2)$
$\frac{1}{5}$	5	2	$T_5^3(5, 3)$

In particular, consider the category with # -1's = 1; there are $\binom{5}{1}$ such sequences. The largest value is $s^+ = 2$ which occurs in exactly one way, namely 1111-1. $c_4^2(2, 4) = 1$. The second largest value is $s^+ = 1$. There are two sequences, namely 111-11 and 1111-1, such that $\frac{1}{3}\sum_{i=1}^{3} x_i - \frac{1}{2}\sum_{i=4}^{5} x_i = 1$. Of these only 111-11 satisfies $S_5^+ = 1$. The other sequence satisfies $S_5^+ = 2$. $c_4^+(1, 3) = 1$. The next largest value is $s^+ = \frac{2}{3}$. There are three sequences, namely 11-111, 111-11 and 1111-1, such that $\frac{1}{2}\sum_{i=1}^{2} x_i - \frac{1}{3}\sum_{i=3}^{5} x_i = \frac{2}{3}$. Of these only 11-111 satisfies $S_5^+ = \frac{2}{3}$. The other two sequences have $S_5^+ > \frac{2}{3}$. $c_4^+(\frac{2}{3}, 2) = 1$. The remaining value is $s^+ = \frac{3}{5}$. All 5 sequences satisfy $\frac{1}{5}\sum_{i=1}^{5} x_i = \frac{3}{5}$. Of these only 1-1111 and -11111 satisfy $S_5^+ = \frac{3}{5}$. The other sequences have $S_5^+ > \frac{3}{5}$. $c_4^+(\frac{3}{5}, 5) = 2$. All five sequences have been counted.

Example 3. For $n = 5$ the $c_k(s, j)$'s are found for # -1's = 0, 1, 2 .

-1's = 0

s	j	$c_5(s, j)$	Types of Sequences
1	0	$\frac{1}{2}$	$T_5^5(0, 0)$
1	5	$\frac{1}{2}$	$T_5^5(5, 5)$

-1's = 1

s	j	$c_4(s, j)$	Types of Sequences
2	4	1	$T_5^4(4, 4)$
2	1	1	$T_5^4(1, 0)$
1	2	1	$T_5^4(2, 1)$
1	3	1	$T_5^4(3, 3)$
$\frac{2}{3}$	2	$\frac{1}{2}$	$T_5^4(2, 2)$
$\frac{2}{3}$	3	$\frac{1}{2}$	$T_5^4(3, 2)$

$\# \text{-1's} = 2$

s	j	$c_3(s, j)$	Types of Sequences
2	2	1	T_5^3 (2, 0)
2	3	1	T_5^3 (3, 3)
$\frac{3}{2}$	1	$\frac{5}{2}$	T_5^3 (1, 0)
$\frac{3}{2}$	4	$\frac{5}{2}$	T_5^3 (4, 3)
$\frac{4}{3}$	3	1	T_5^3 (3, 1)
$\frac{4}{3}$	2	1	T_5^3 (2, 2)
1	1	$\frac{1}{2}$	T_5^3 (1, 1)
1	4	$\frac{1}{2}$	T_5^3 (4, 2)

Consider the category with $\#$ -1's = 1. The largest value is $s = 2$ which occurs in exactly 2 ways, namely for 1111-1 and -1111; $c_4(2, 4) = c_4(2, 1) = 1$. The second largest value is $s = 1$. There are four sequences satisfying $\left| \frac{1}{2} \sum_{i=1}^{2} x_i - \frac{1}{3} \sum_{i=3}^{5} x_i \right| = 1$ or $\left| \frac{1}{3} \sum_{i=1}^{3} x_i - \frac{1}{2} \sum_{i=4}^{5} x_i \right| = 1$; they are -11111, 1-1111, 111-11, 1111-1. Of these only 1-1111 and 111-11 satisfy $S_5 = 1$. The other sequences have $S_5 = 2$. $c_4(1, 2) = c_4(1, 3) = 1$. The third largest value is $s = \frac{2}{3}$. All 5 sequences satisfy $\left| \frac{1}{2} \sum_{i=1}^{2} x_i - \frac{1}{3} \sum_{i=3}^{5} x_i \right| = \frac{2}{3}$ or $\left| \frac{1}{3} \sum_{i=1}^{3} x_i - \frac{1}{2} \sum_{i=4}^{5} x_i \right| = \frac{2}{3}$. Of these only 11-111 satisfies $S_5 = \frac{2}{3}$. The

other sequences have $S_5 > \frac{2}{3}$. $J_5(11\text{-}111) = 2, 3$ so $R_5(11\text{-}111) =$
$\begin{cases} 2 & \text{w.p. } \frac{1}{2} \\ 3 & \text{w.p. } \frac{1}{2} \end{cases}$. Hence $c_4(\frac{2}{3}, 2) = c_4(\frac{2}{3}, 3) = \frac{1}{2}$. All five sequences have been counted.

The joint distributions (S_n^+, R_n^+) and (S_n, R_n) are available from the lists constructed in III. The $c_k^+(s^+, j)$'s and $c_k(s, j)$'s are used to find $P((S_n^+, R_n^+) = (s^+, j) \mid p)$ and $P((S_n^+, R_n^+) = (s^+, j) \mid p)$ according to (2.1.1).

IV. The marginal distributions S_n^+ and S_n are found from the joint distributions in III. according to (2.1.2). Since these distributions will be used to find critical regions for the tests H_0 vs. H_1^+ and H_0 vs. H_1, the s^+ and s values must be ordered over all categories of # -1's. The terms for # -1's $> \left[\frac{n}{2}\right]$ must be included.

Example 4. The marginal distributions S_5^+ and S_5 are found.

s^+	$P(S_5^+ = s^+ \mid p)$
2	$p(1-p)^4 + p^4(1-p) + p^2(1-p)^3 + p^3(1-p)^2$
$\frac{3}{2}$	$3p^2(1-p)^3 + 3p^3(1-p)^2$
$\frac{4}{3}$	$p^2(1-p)^3 + p^3(1-p)^2$
1	$(1-p)^5 + p^5 + p(1-p)^4 + p^4(1-p) + 2p^2(1-p)^3 + 2p^3(1-p)^2$

$\frac{2}{3}$	$p(1-p)^4 + p^4(1-p)$
$\frac{3}{5}$	$2p(1-p)^4 + 2p^4(1-p)$
$\frac{1}{3}$	$p^2(1-p)^3 + p^3(1-p)^2$
$\frac{1}{5}$	$2p^2(1-p)^3 + 2p^3(1-p)^2$

s	$P(S_5 = s \mid p)$
2	$2p(1-p)^4 + 2p^4(1-p) + 2p^2(1-p)^3 + 2p^3(1-p)^2$
$\frac{3}{2}$	$5p^2(1-p)^3 + 5p^3(1-p)^2$
$\frac{4}{3}$	$2p^2(1-p)^3 + 2p^3(1-p)^2$
1	$(1-p)^5 + p^5 + 2p(1-p)^4 + 2p^4(1-p)$ $+ p^2(1-p)^3 + p^3(1-p)^2$
$\frac{2}{3}$	$p(1-p)^4 + p^4(1-p)$

For given p the critical regions for a size-α test may now be computed. By Theorems 2 and 6 the critical regions need only be computed for $p \leq \frac{1}{2}$.

V. Lastly, the marginal distributions R_n^+ and R_n are found from the joint distributions in III. according to (2.1.4). The terms for # -1's $> \left[\frac{n}{2}\right]$ must be included.

Example 5. The marginal distributions R_5^+ and R_5 are found.

j	$P(R_5^+ = j \mid p)$
0	$(1-p)^5 + 2p(1-p)^4 + 2p^2(1-p)^3$
1	$(1-p)^4 + 3p^2(1-p)^3 + 2p^3(1-p)^2$
2	$p(1-p)^4 + p^4(1-p) + 2p^2(1-p)^3 + p^3(1-p)^2$
3	$p^4(1-p) + p(1-p)^4 + 2p^3(1-p)^2 + p^2(1-p)^3$
4	$p^4(1-p) + 3p^3(1-p)^2 + 2p^2(1-p)^3$
5	$p^5 + 2p^4(1-p) + 2p^3(1-p)^2$

j	$P(R_5 = j \mid p)$
0	$\frac{1}{2} p^5 + \frac{1}{2}(1-p)^5$
1	$p^4(1-p) + p(1-p)^4 + 3p^3(1-p)^2 + 3p^2(1-p)^3$
2	$\frac{3}{2} p^4(1-p) + \frac{3}{2} p)1-p)^4 + 2p^3(1-p)^2 + 2p^2(1-p)^3$
3	$\frac{3}{2} p^4(1-p) + \frac{3}{2} p(1-p)^4 + 2p^3(1-p)^2 + 2p^2(1-p)^3$
4	$p^4(1-p) + p(1-p)^4 + 3p^3(1-p)^2 + 3p^2(1-p)^3$
5	$\frac{1}{2} p^5 + \frac{1}{2}(1-p)^5$

One need only compute the probabilities for $R_n^+ = 0, 1, \cdots, \left[\frac{n}{2}\right]$ and $p \epsilon (0, 1)$ or $R_n^+ = 0, 1, \cdots, n$ and $p \epsilon (0, \frac{1}{2}]$; see Theorem 2. Also one need only compute the probabilities for $p \leq \frac{1}{2}$ and $R_n = 0, 1, \cdots, \left[\frac{n}{2}\right]$; see Theorem 6.

2.3 Properties of S_n^+

Properties of S_n^+ when the null hypothesis is true are studied. Included first is a result for reducing the calculations needed to evaluate S_n^+ for a sample sequence. An analysis of $o(\Omega_{S_n^+})$ follows. The behavior of the probability density and its mode, mean, variance and median are investigated for varying n and p; critical regions for the test H_0 vs. H_1^+ are tabulated and $P(S_n^+ \geq s^+ \mid p)$ is studied as a function of p, n and s^+. Indications for testing when $n>20$ are discussed.

Theorem 10 is useful for reducing the calculations needed to evaluate S_n^+ in practice. First we prove Lemma 9.

<u>Lemma 9.</u> If $S_n^+(X_1, \cdots, X_n) = \frac{1}{j} \sum_{i=1}^{j} X_i - \frac{1}{n-j} \sum_{i=j+1}^{n} X_i$ where $j \neq 0, n$, then $\frac{1}{j-1} \sum_{i=1}^{j-1} X_i < 1$ or $\frac{1}{n+1-j} \sum_{i=j}^{n} X_i > -1$ when $j>1$ and $\frac{1}{j+1} \sum_{i=1}^{j+1} X_i < 1$ or $\frac{1}{n-1-j} \sum_{i=j+2}^{n} X_i > -1$ when $j<n-1$. If $j = 1$ then $\frac{1}{n} \sum_{i=1}^{n} X_i > -1$ and if $j = n-1$ then $\frac{1}{n} \sum_{i=1}^{n} X_i < 1$.

<u>Proof.</u> Suppose $S_n^+(X_1, \cdots, X_n) = \frac{1}{j} \sum_{i=1}^{j} X_i - \frac{1}{n-j} \sum_{i=j+1}^{n} X_i$ and $j>1$. If $\frac{1}{j-1} \sum_{i=1}^{j-1} X_i = 1$ and $\frac{1}{n+1-j} \sum_{i=j}^{n} X_i = -1$ then

$2 = \frac{1}{j-1} \sum_{i=1}^{j-1} X_i - \frac{1}{n+1-j} \sum_{i=j}^{n} X_i$; this implies that $\frac{1}{j} \sum_{i=1}^{j} X_i - \frac{1}{n-j} \sum_{i=j+1}^{n} X_i = \frac{1}{j}(j-1-1) - \frac{1}{n-j}(-(n-j)) = 2 - \frac{2}{j} < 2$, a contradiction.

A similar argument is used to verify the remaining statements of the theorem.

■

<u>Theorem 10.</u> If $S_n^+(X_1, \cdots, X_n) = \frac{1}{j} \sum_{i=1}^{j} X_i - \frac{1}{n-j} \sum_{i=j+1}^{n} X_i$ and $j \neq 0, n$, then $X_j = 1$ and $X_{j+1} = -1$.

<u>Proof.</u> Suppose $j \neq 0, 1, n$ and $X_j = -1$, then $\sum_{j=1}^{j} X_i = \sum_{i=1}^{j-1} X_i - 1$ and $\sum_{i=j+1}^{n} X_i = \sum_{i=j}^{n} X_i + 1$ so that

$$\frac{1}{j} \sum_{i=1}^{j} X_i - \frac{1}{n-j} \sum_{i=j+1}^{n} X_i = \frac{1}{j} \left(\sum_{i=1}^{j-1} X_i - 1 \right) - \frac{1}{n-j} \left(\sum_{i=j}^{n} X_i + 1 \right)$$

$$= \frac{1}{j} \sum_{i=1}^{j-1} X_i - \frac{1}{n-j} \sum_{i=j}^{n} X_i - \frac{1}{j} - \frac{1}{n-j}$$

$$= \frac{1}{j-1} \sum_{i=1}^{j-1} X_i - \frac{1}{n-j+1} \sum_{i=j}^{n} X_i$$

$$- \frac{1}{j} \left(\frac{1}{j-1} \sum_{i=1}^{j-1} X_i + 1 \right) - \frac{1}{n-j} \left(\frac{1}{n-j+1} \sum_{i=j}^{n} X_i + 1 \right)$$

$$< \frac{1}{j-1} \sum_{i=1}^{j-1} X_i - \frac{1}{n-j+1} \sum_{i=j}^{n} X_i$$

since either $\frac{1}{j-1} \sum_{i=1}^{n-1} X_i < 1$ or $\frac{1}{n-j+1} \sum_{i=j}^{n} X_i > -1$ by Lemma 9.

Thus for $j \neq 0, 1, n$, if $X_j = -1$ then $R_n^+(X_1, \cdots, X_n) \neq j$, a contradiction. Suppose $j = 1$ and $X_1 = -1$; then

$$\frac{1}{j} \sum_{i=1}^{j} X_i - \frac{1}{n-j} \sum_{i=j+1}^{n} X_i = -1 - \frac{1}{n-1} \sum_{i=2}^{n} X_i \geq 0$$

$$\Rightarrow \frac{1}{n-1} \sum_{i=2}^{n} X_i = -1 \quad \Rightarrow X_2 = X_3 = \cdots = X_n = -1 .$$

But $X_1, \cdots, X_n = -1, \cdots, -1$ has $S_n^+ = 1$ and $R_n^+ = 0$, a contradiction. Hence, if $R_n^+ (X_1, \cdots, X_n) = j$ where $j \neq 0, n$ then $X_j = 1$.

A similar argument shows that $X_{j+1} = -1$. ■

Hence to evaluate S_n^+ for a sample sequence, one need only calculate $\frac{1}{r} \sum_{i=1}^{r} x_i - \frac{1}{n-r} \sum_{i=r+1}^{n} x_i$ for those $r \in \{1, \cdots, n-1\}$ where $x_r = 1$ and $x_{r+1} = -1$, and if each value calculated is less than 1 also for $r = 0$ if the number of 1's in $x_1, \cdots, x_n$ is less than or equal to the number of -1's or for $r = n$ otherwise.

No formula for the order of the sample space S_n^+ has been found, but it is of interest to study the behavior of $o(\Omega_{S_n^+})$ as a function of n. $o(\Omega_{S_n^+})$ vs. n is plotted in Figure 3 for $n = 1$-$10, 15, 20$. Note that $o(\Omega_{S_n^+})$ increases rapidly as n increases. This behavior is expected: Consider the lists constructed in I. of the procedure for counting. For given r and for each a the number of types of sequences $T_n^k(r, a)$ which have $\frac{1}{r} \sum_{i=1}^{r} X_i - \frac{1}{n-r} \sum_{i=r+1}^{n} X_i \geq \left[\frac{n-2k}{n}\right]$ increases with n. Moreover the number of r values with $r \leq \left[\frac{n}{2}\right]$ increases so that the collection of lists grows. Also as r increases so does the number of types of sequences with a 1's in the first r places. Many of the values obtained using the procedure for given n are unique.

$\log_2(o(\Omega_{S_n^+}))$ vs. $\log_2 n$ is plotted in Figure 4 ; note that for 5-10, 15, 20 the graph is nearly a straight line. The best least squares straight line fitting these points is $\log_2(o(\Omega_{S_n^+})) = 2.296 \log_2 n - 2.456$, so an approximation to $o(\Omega_{S_n^+})$ may be found by $o(\Omega_{S_n^+}) = 2^{-2.456} \cdot n^{2.296}$.

Next we investigate the behavior of the probability density S_n^+ and its mode, mean, variance and median for varying n and p. In Figures 5.1, 5.2 are graphed $P(S_n^+ = s^+ \mid p)$ vs. s^+ for $n = 10$ and 20, for $p = .1, .2, .3, .4, .5$ and for those s^+ for which $P(S_n^+ = s^+ \mid p) \geq .01$. The probability densities for $p = .6, .7, .8, .9$ are the same as those for $1 - p$ by Theorem 2 and hence are omitted.

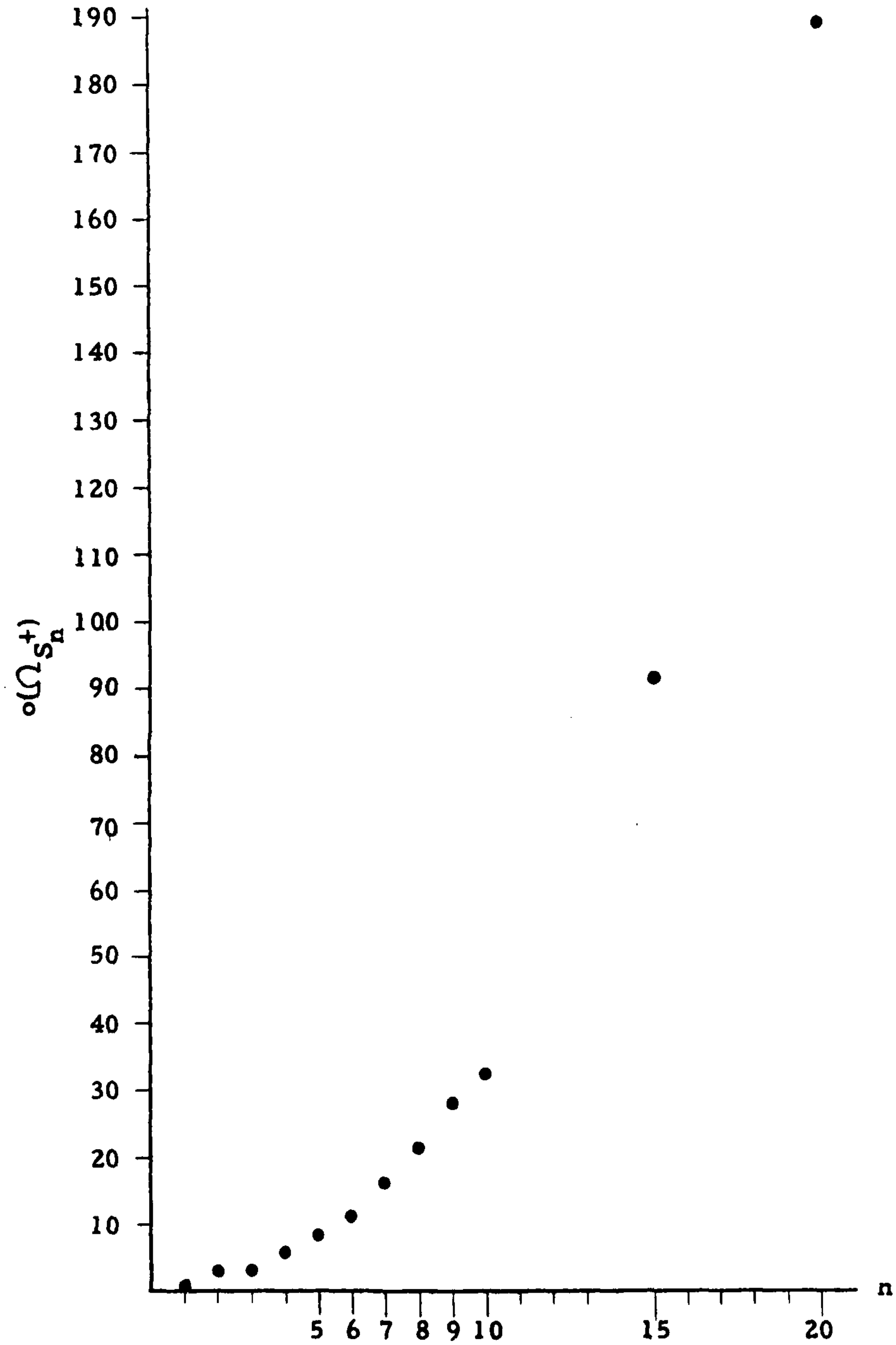


Figure 3. $o(\Omega_{S_n^+})$ vs. n.

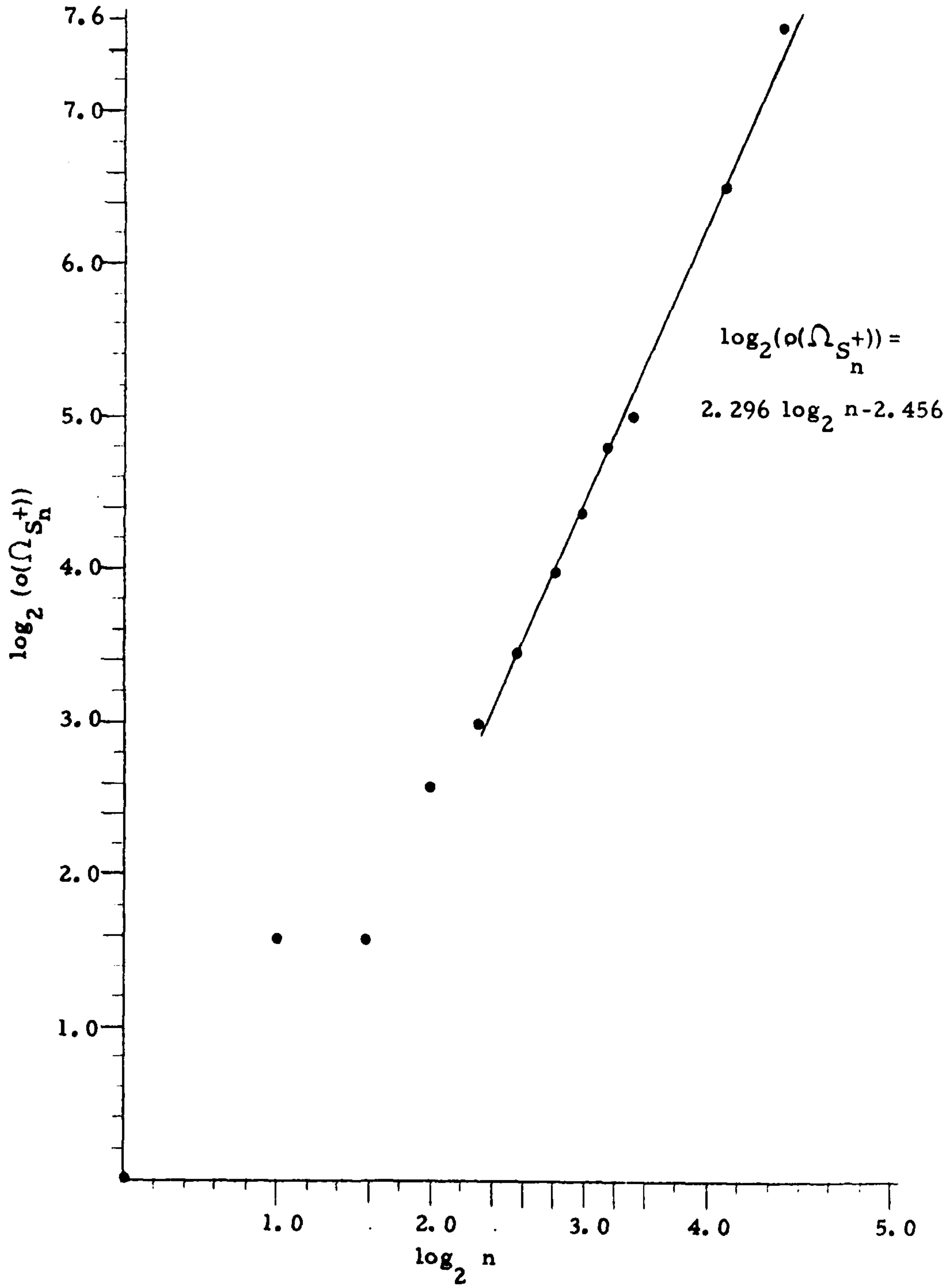


Figure 4. $\log_2 (o(\Omega_{S_n^+}))$ vs. $\log_2 n$.

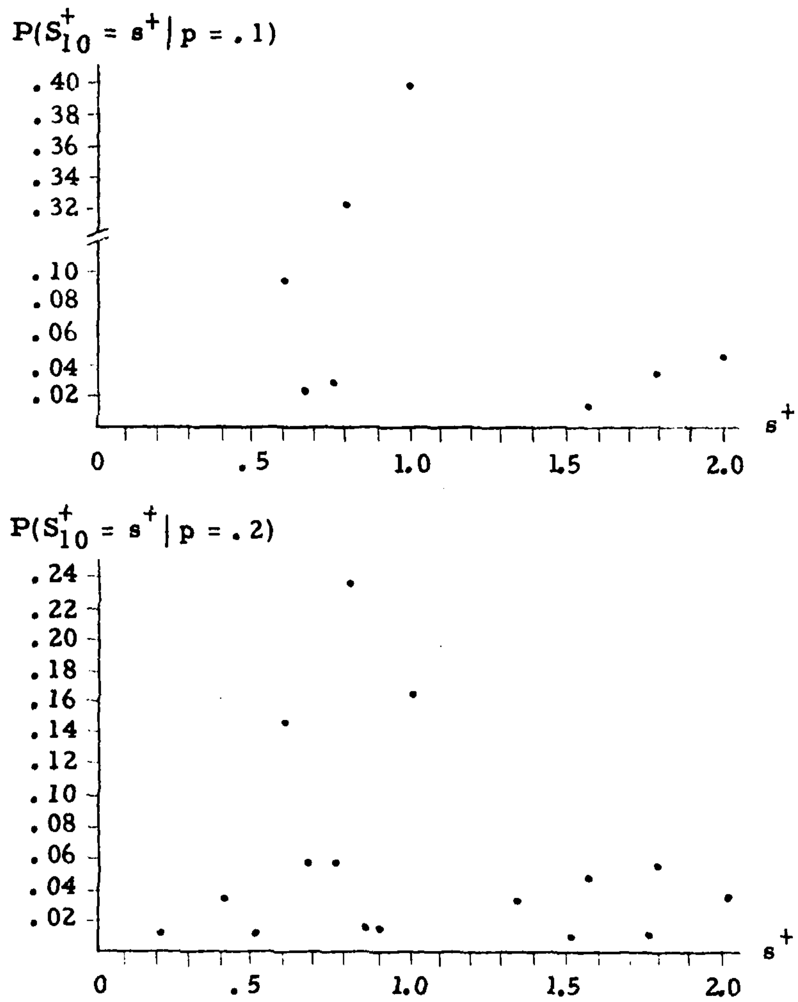


Figure 5.1 $P(S_{10}^{+} = s^{+} | p)$ vs. s^{+}.

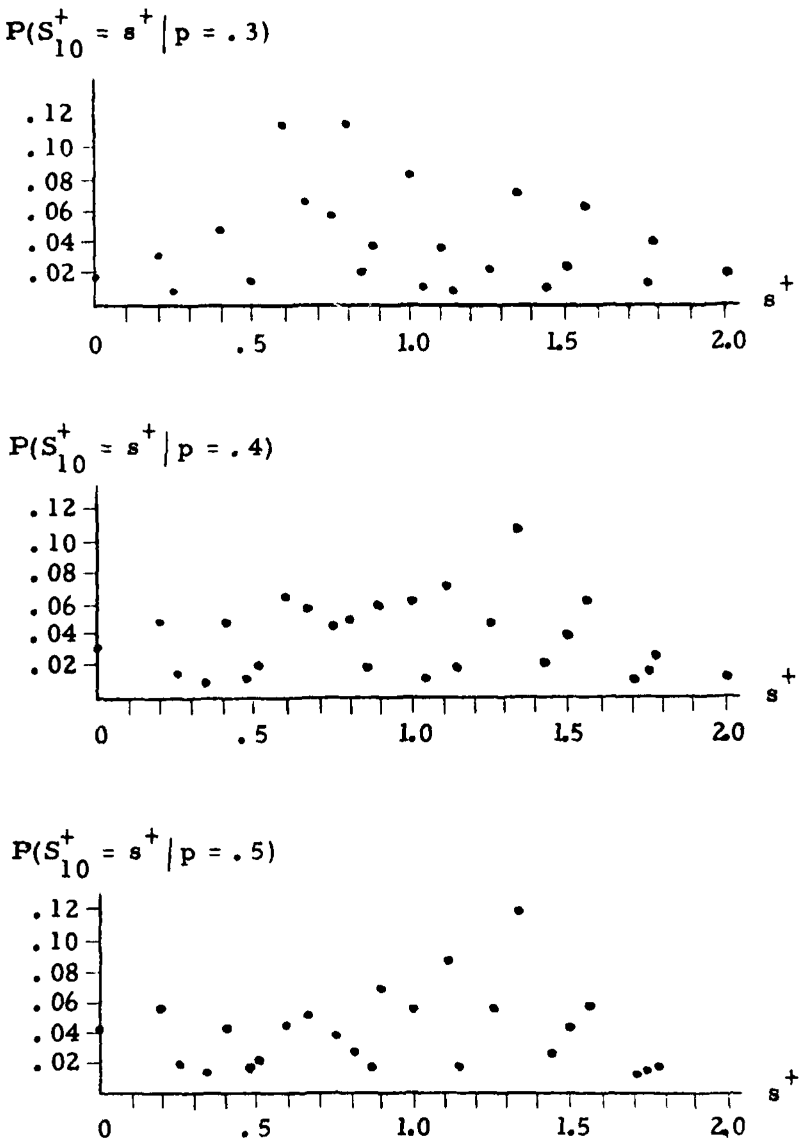


Figure 5.1 (Continued).

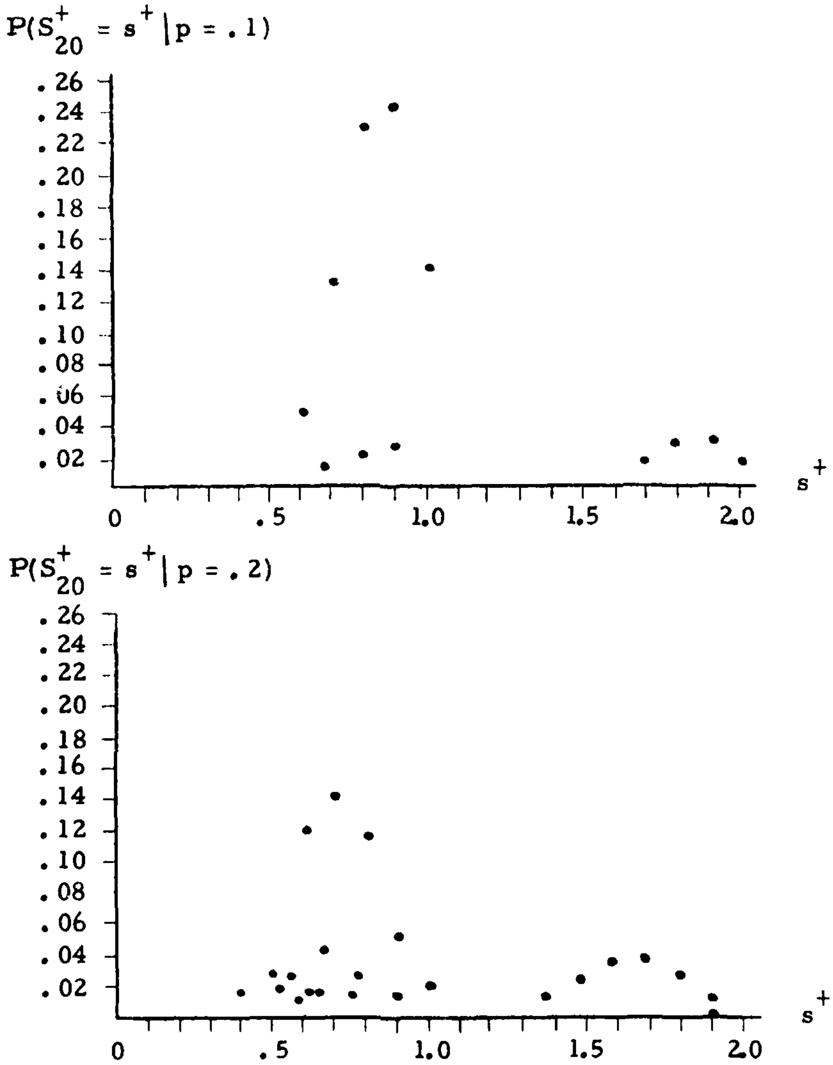


Figure 5.2 $P(S_{20}^{+} = s^{+} | p)$ vs. s^{+}.

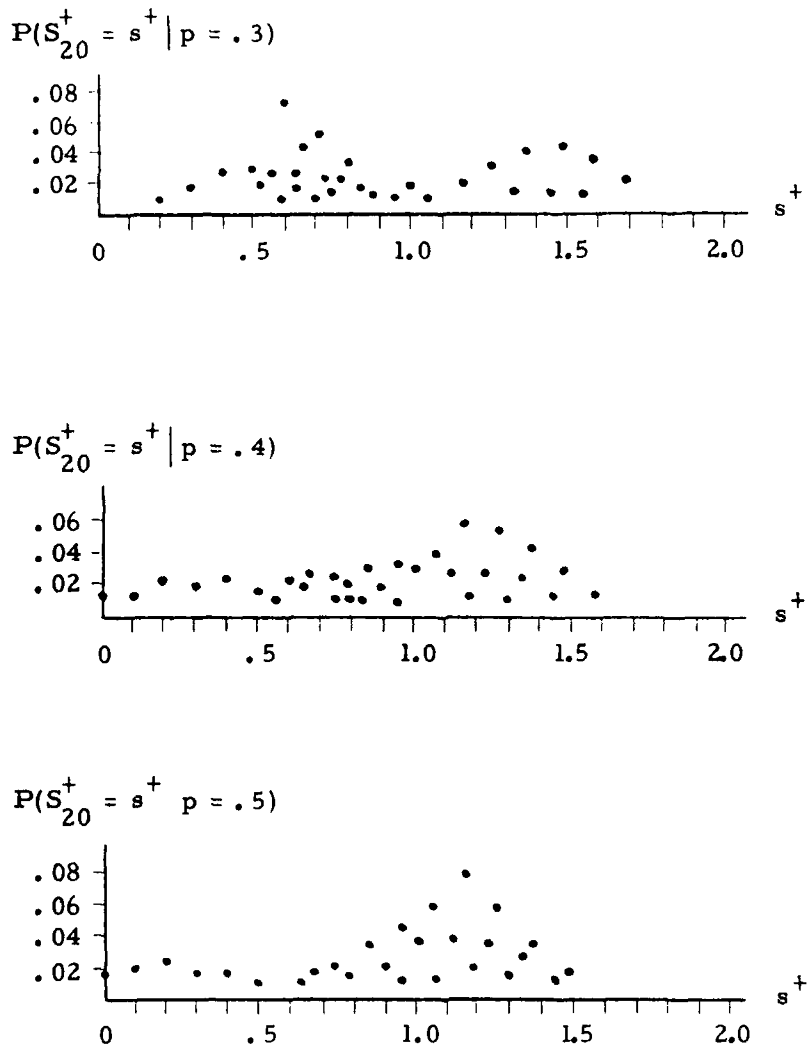


Figure 5. 2 (Continued).

Only the "large" probability masses are included since $o(\Omega_{S_n^+})$ is very large.

Note that for each n plotted as p: $.1 \longrightarrow .5$ the number of "large" probability masses for s^+ near 2 decreases and for s^+ near 0 increases ; the s^+ values with "large" probabilities are shifting downwards.

Let $m^+_{n,p} = P(S^+_n = s^+_m \mid p)$ where s^+_m is the mode of S^+_n for those values of n and p. Note that $\sup_p m^+_{n,p} = m^+_{n,0} = 1$ and the mode occurs at $s^+_m = 1$; the only sequence with positive probability when $p = 0$ is the sequence with $x_i = -1 \forall i$. When p is near 0 the mode occurs at $s^+ = 1$ and p must be nearer to 0 in order for the mode to occur at $s^+ = 1$ as n increases. Note the similarity in the shift of the mode as p: $0 \longrightarrow \frac{1}{2}$ for each n. As p increases from 0, $m^+_{n,p}$ decreases and the mode shifts to values $s^+_m < 1$; then as p increases to $\frac{1}{2}$, $m^+_{n,p}$ increases and the mode shifts to values $s^+_m > 1$.

For $n = 5\text{-}10, 15, 20$, $E(S^+_n \mid p)$ vs. p is sketched in Figure 6. The mean is symmetric in p and $1\text{-}p$ because of the symmetry in S^+_n ; $E(S^+_n \mid p{=}0) = 1$ for all n.

For each n plotted $E(S^+_n \mid p)$ decreases as p increases from 0 and then increases as p nears $\frac{1}{2}$. A similar behavior was noted for the mode. The minimums tend farther from 0 and 1 as n

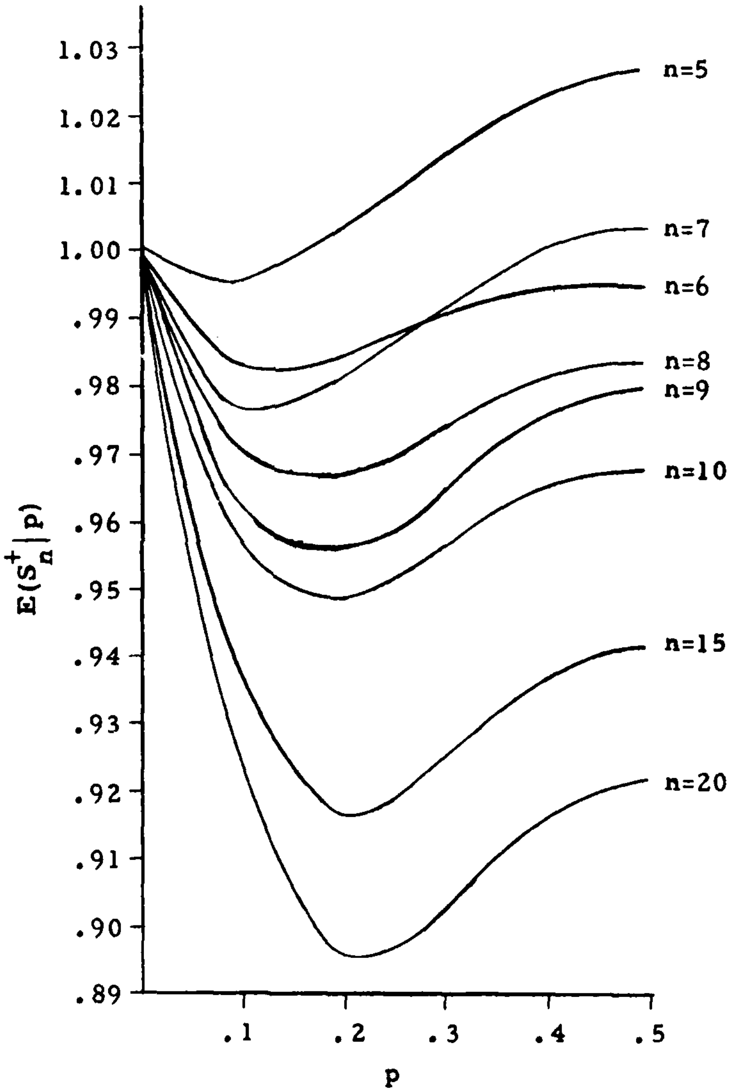


Figure 6. $E(S_n^+|p)$ vs. p .

increases and a local maximum occurs at $p = \frac{1}{2}$ for all n. In general as n increases $E(S_n^+ \mid p)$ decreases for each p.

Note that the range of $E(S_n^+ \mid p)$ over n and p is rather small.

The mean tends to be flatter for n even than for n odd.

Var $(S_n^+ \mid p)$ vs. p is plotted in Figure 7 for $n = 5$-10, 15, 20. Again symmetries in p and $1-p$ occur and $\mathrm{Var}(S_n^+ \mid p=0) = 0$ for all n. For each p, $\mathrm{Var}(S_n^+ \mid p)$ decreases as n increases. For $n = 15, 20$ the variance is nearly constant for $.2 \leq p \leq .8$.

Figure 8 illustrates the behavior of $\mathrm{med}(S_n^+ \mid p)$ for $n = 5, 10, 15, 20$ as p varies. $\mathrm{med}(S_n^+ \mid p) = \mathrm{med}(S_n^+ \mid 1-p)$ for all n and p and $\mathrm{med}(S_n^+ \mid p=0) = 1$ for all n. S_n^+ is a discrete random variable so the median is a discontinuous function of p. We expect that as p changes to $p \pm \delta$ where δ is small that $\mathrm{med}(S_n^+ \mid p)$ and $\mathrm{med}(S_n^+ \mid p \pm \delta)$ will occur at values of s^+ near each other, however, since $P(S_n^+ \geq s^+ \mid p)$ is of the form $\sum_{k=0}^{n} C_k(s^+)\, p^k (1-p)^{n-k}$.

The median as a function of p behaves as does the mode and mean: the median decreases as p increases from 0 and then increases as p nears $\frac{1}{2}$. The following theorem and conjecture suggest the median has a local maximum at $p = \frac{1}{2}$ for all n.

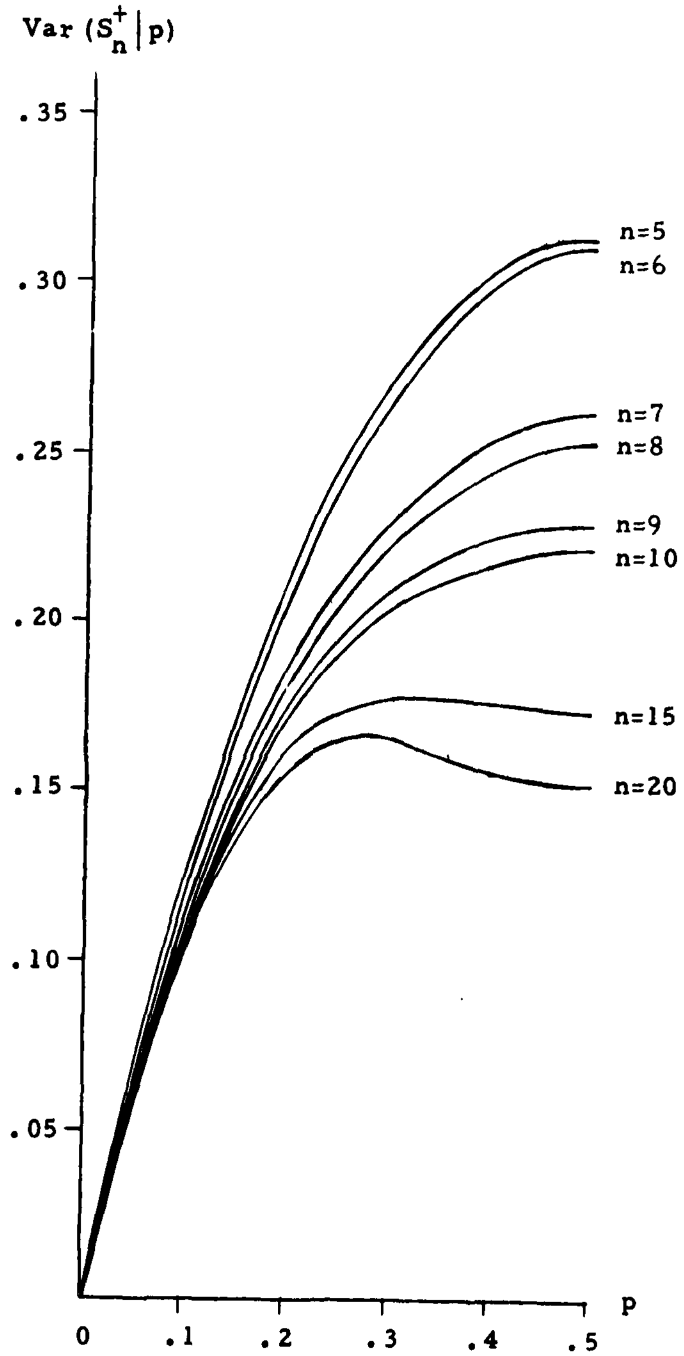


Figure 7. Var $(S_n^+ | p)$ vs. p .

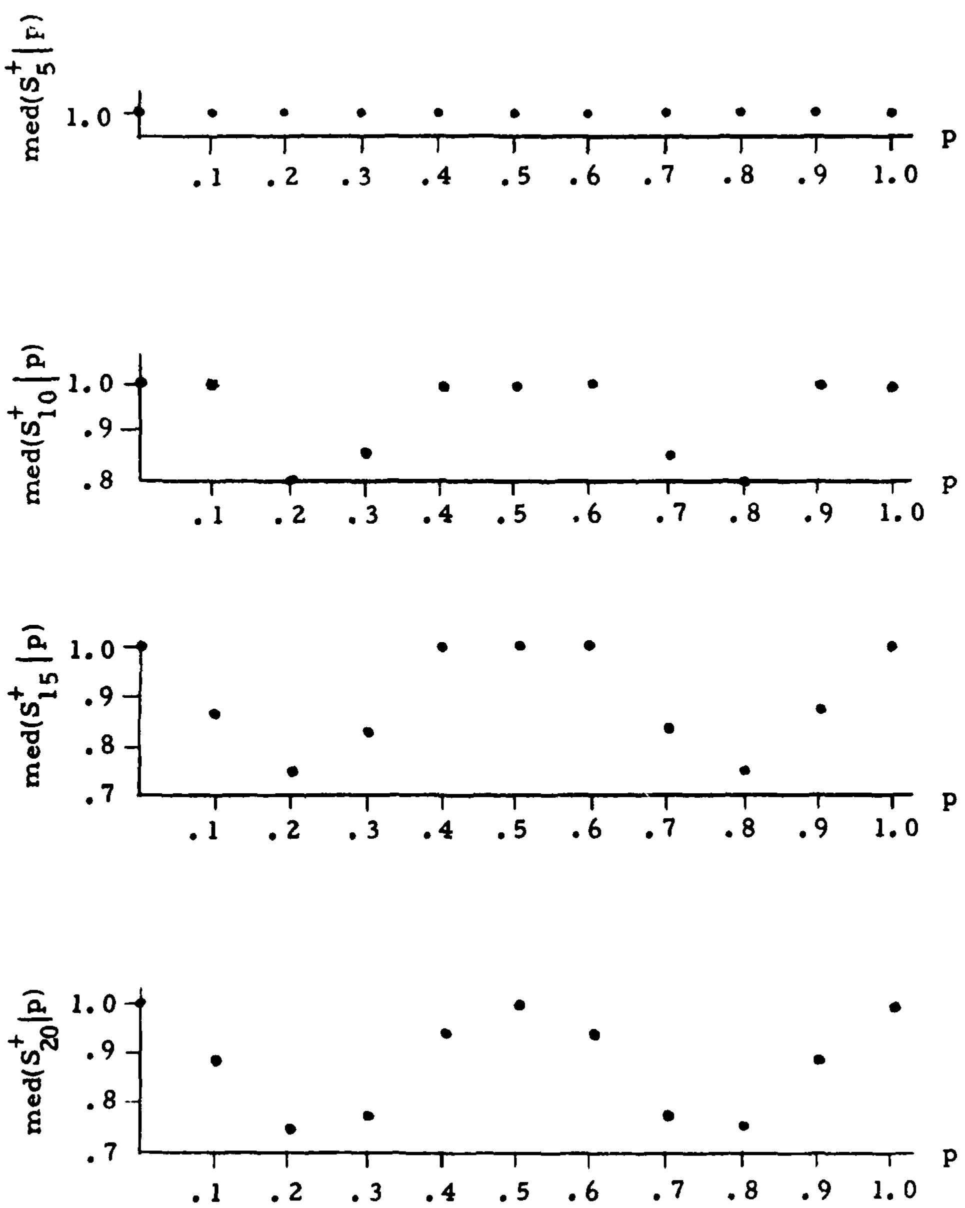


Figure 8. med $(S_n^+ | p)$ vs. p .

<u>Theorem 11.</u> For all n, $\operatorname{med}(S_n^+ \mid p=\frac{1}{2}) > 1$.

<u>Proof.</u> We show that the number of sequences with $S_n^+ \geq 1$ is at least 2^{n-1}. Note the following.

(i) The sequence with n 1's and the sequence with n -1's each have $S_n^+ = 1$.

(ii) Any sequence with k -1's where $k \leq \left[\frac{n}{2}\right]$ and $x_n = -1$ has $S_n^+ \geq \frac{n-1-2(k-1)}{n-1} + 1 \geq 1$.

(iii) Any sequence with k -1's where $k \geq \left[\frac{n}{2}\right] + 1$ and $x_1 = 1$ has $S_n^+ \geq 1 - \frac{n-1-2k}{n-1} \geq 1$.

(iv) For $1 \leq k \leq \left[\frac{n-1}{2}\right]$ the sequence with $n-2k$ 1's in the first $n-2k$ places, k -1's in the next $2k-1$ places and a 1 in the last place each have $S_n^+ \geq 1 - \frac{k-k}{2k} = 1$.

(v) For $1 \leq k \leq \left[\frac{n-1}{2}\right]$ the sequences with a -1 in the first place, k 1's in the next $2k-1$ places and $n-2k$ -1's in the last $n-2k$ places each have $S_n^+ \geq \frac{k-k}{2k} - (-1) = 1$.

(vi) For $k = \frac{n}{2}$ the sequence with a 1 in the first place, $\frac{n}{2}$ -1's in the next $n-2$ places and a 1 in the last place each have $S_n^+ \geq 1$.

All of the sequences described above are unique; the total number of these sequences is Q where

$$Q = 2 + \sum_{k=0}^{\left[\frac{n-2}{2}\right]} \binom{n-1}{k} + \sum_{k=\left[\frac{n+2}{2}\right]}^{n-1} \binom{n-1}{k} + 2 \sum_{k=1}^{\left[\frac{n-1}{2}\right]} \binom{2k-1}{k} + Q_0$$

and $Q_0 = \begin{cases} 0 & \text{if } n \text{ is odd} \\ \binom{n-2}{\frac{n}{2}} & \text{if } n \text{ is even} \end{cases}$.

$$\text{Now } Q = 2 + \sum_{k=0}^{n-1} \binom{n-1}{k} - \binom{n-1}{\left[\frac{n}{2}\right]} + 2 \sum_{k=1}^{\left[\frac{n-1}{2}\right]} \binom{2k-1}{k} + Q_0$$

$$\geq \sum_{k=0}^{n-1} \binom{n-1}{k} = 2^{n-1}$$

since $2 + 2 \sum_{k=1}^{\left[\frac{n-1}{2}\right]} \binom{2k-1}{k} + Q_0 \geq \binom{n-1}{\left[\frac{n}{2}\right]}$. If n is even we have

$$2 + 2 \sum_{k=1}^{\frac{n-2}{2}} \binom{2k-1}{k} + \binom{n-2}{\frac{n}{2}} \geq 2 \binom{2(\frac{n-2}{2})-1}{\frac{n-2}{2}} + \binom{n-2}{\frac{n}{2}}$$

$$= 2 \binom{n-3}{\frac{n-2}{2}} + \binom{n-2}{\frac{n}{2}}$$

$$= \binom{n-3}{\frac{n-4}{2}} + \binom{n-3}{\frac{n-2}{2}} + \binom{n-2}{\frac{n}{2}}$$

$$= \binom{n-2}{\frac{n-2}{2}} + \binom{n-2}{\frac{n}{2}}$$

$$= \binom{n-1}{\frac{n}{2}} \quad .$$

If n is odd we have

$$2 + 2 \sum_{k=1}^{\frac{n-1}{2}} \binom{2k-1}{k} \geq 2 \binom{2(\frac{n-1}{2})-1}{\frac{n-1}{2}} = 2 \binom{n-2}{\frac{n-1}{2}}$$

$$= \binom{n-2}{\frac{n-3}{2}} + \binom{n-2}{\frac{n-1}{2}} = \binom{n-1}{\frac{n-1}{2}} \quad . \quad \blacksquare$$

Note that all sequences with $S_n^+ \geq 1$ have not been included in the counting for Q.

<u>Conjecture</u>. For all n and p, $\mathrm{med}(S_n^+ \mid p) \leq 1$.

The proof of the conjecture requires one to first count <u>all</u> those sequences with $S_n^+ \geq 1$ for each n. Such a task appears to be impossible.

As a consequence of Theorem 11, note that $E(S_n^+ \mid p=\frac{1}{2}) \geq \frac{1}{2}$ for all n.

Critical regions for the test H_0 vs. H_1^+ are contained in Table 9 for $n = 5, 10, 15, 20$ and $p = .1, .2, .3, .4, .5$. Critical regions for $p = .6, .7, .8, .9$ are also available from the table. See Theorem 2.

Table 9. $P(S_n^+ \geq s^+ \mid p)$.

n=5

s^+ \ p	.1	.2	.3	.4	.5
2/1	.074	.109	.122	.125	.125
3/2	.098	.186	.254	.298	.313
4/3	.106	.211	.298	.355	.375

n=6

s^+ \ p	.1	.2	.3	.4	.5
2/1	.066	.087	.087	.081	.078
8/5	.093	.157	.189	.201	.203
4/3	.103	.195	.261	.300	.313
6/5	.109	.228	.335	.411	.438

n=7

s^+ \ p	.1	.2	.3	.4	.5
2/1	.060	.070	.061	.051	.047
5/3	.089	.136	.154	.132	.125
3/2	.094	.161	.199	.215	.219
6/5	.108	.223	.326	.397	.422

n=8

s^+ \ p	.1	.2	.3	.4	.5
2/1	.054	.056	.043	.032	.027
12/7	.086	.119	.109	.085	.074
5/3	.089	.133	.136	.121	.113
3/2	.090	.143	.160	.159	.156
10/7	.099	.182	.235	.259	.266
8/7	.109	.229	.344	.429	.461

Table 9 (Continued).

n=9

s^+ \ p	.1	.2	.3	.4	.5
2/1	.048	.045	.030	.019	.016
14/8	.082	.104	.083	.055	.043
12/7	.085	.116	.103	.078	.066
5/3	.085	.120	.113	.095	.086
8/5	.086	.122	.119	.105	.098
3/2	.097	.168	.194	.190	.184
4/3	.104	.191	.243	.265	.270
11/10	.108	.225	.337	.419	.449

n=10

s^+ \ p	.1	.2	.3	.4	.5
2/1	.044	.036	.021	.012	.009
16/9	.078	.089	.063	.035	.024
14/8	.081	.101	.079	.050	.038
12/7	.082	.104	.086	.061	.050
5/3	.082	.105	.089	.068	.059
8/5	.082	.106	.093	.073	.064
14/9	.095	.153	.157	.133	.119
4/3	.105	.200	.264	.295	.303
22/21	.110	.234	.358	.452	.487

n=15

s^+ \ p	.1	.2	.3	.4	.5
2/1	.026	.012	.004	.001	.000
26/14	.059	.040	.015	.004	.001
14/8	.063	.049	.022	.008	.005
12/7	.085	.092	.052	.020	.010
16/10	.087	.104	.069	.034	.022
11/7	.096	.143	.114	.062	.040
20/13	.097	.150	.130	.079	.055
41/28	.097	.153	.137	.090	.067
13/9	.097	.153	.139	.096	.075
20/14	.100	.178	.189	.147	.119
14/13	.108	.221	.335	.431	.470

Table 9 (Continued).

n=20

s^+ \ p	.1	.2	.3	.4	.5
2/1	.015	.004	.001	.000	.000
36/19	.042	.017	.003	.000	.000
9/5	.045	.021	.005	.001	.000
34/19	.071	.048	.015	.003	.001
22/13	.074	.056	.021	.005	.002
32/19	.089	.093	.043	.011	.003
5/3	.090	.101	.050	.014	.005
38/24	.090	.103	.055	.018	.008
30/19	.097	.137	.090	.033	.014
3/2	.097	.146	.107	.049	.026
28/19	.099	.170	.150	.077	.042
16/11	.099	.170	.150	.079	.044
26/18	.100	.174	.164	.093	.056
24/17	.100	.175	.168	.102	.066
136/99	.100	.176	.171	.108	.074
26/19	.100	.189	.210	.151	.109
4/3	.102	.193	.223	.175	.137
100/99	.109	.220	.332	.442	.493

Since $P(S_n^+ \geq s^+ \mid p)$ is a continuous function of p, approximations to critical regions where p assumes a value other than those listed can be made. For example, suppose $n = 15$, $p = .35$ and $S_n^+ = \frac{22}{13}$. Plotting $P(S_{15}^+ \geq \frac{22}{13} \mid p)$ vs. p we find that $P(S_{15}^+ \geq \frac{22}{13} \mid p = .35) = .04$; see Figure 10. $P(S_n^+ \geq s^+ \mid p)$ is not a continuous function in s^+, however. Hence interpolation for a significance test between two s^+ values in the table is hazardous. The values attained by S_n^+ are not uniformly spaced throughout any interval in critical regions nor are the probabilities of the values of equal size.

The following interesting properties (A) and (B) of $P(S_n^+ \geq s^+ \mid p)$ are suggested from computations.

(A) For $s^+ > 1$ and any n, $P(S_n^+ \geq s^+ \mid p) \begin{cases} = 0 \text{ for } p = 0, 1 \\ > 0 \text{ for } p \epsilon (0, 1) \end{cases}$

and either (A1) two local maximums occur at say p' and $1-p'$ where $p' \neq \frac{1}{2}$ and a local minimum occurs at $p = \frac{1}{2}$ or (A2) a maximum occurs at $p = \frac{1}{2}$.

(B) For $s^+ \leq 1$ and any n, $P(S_n^+ \geq s^+ \mid p) \begin{cases} = 1 \text{ for } p = 0, 1 \\ < 1 \text{ for } p \epsilon (0, 1) \end{cases}$

and either (B1) two local minimums occur at say p' and $1-p'$ where $p' \neq \frac{1}{2}$ and a local maximum occurs at $p = \frac{1}{2}$ or (B2) a minimum occurs at $p = \frac{1}{2}$.

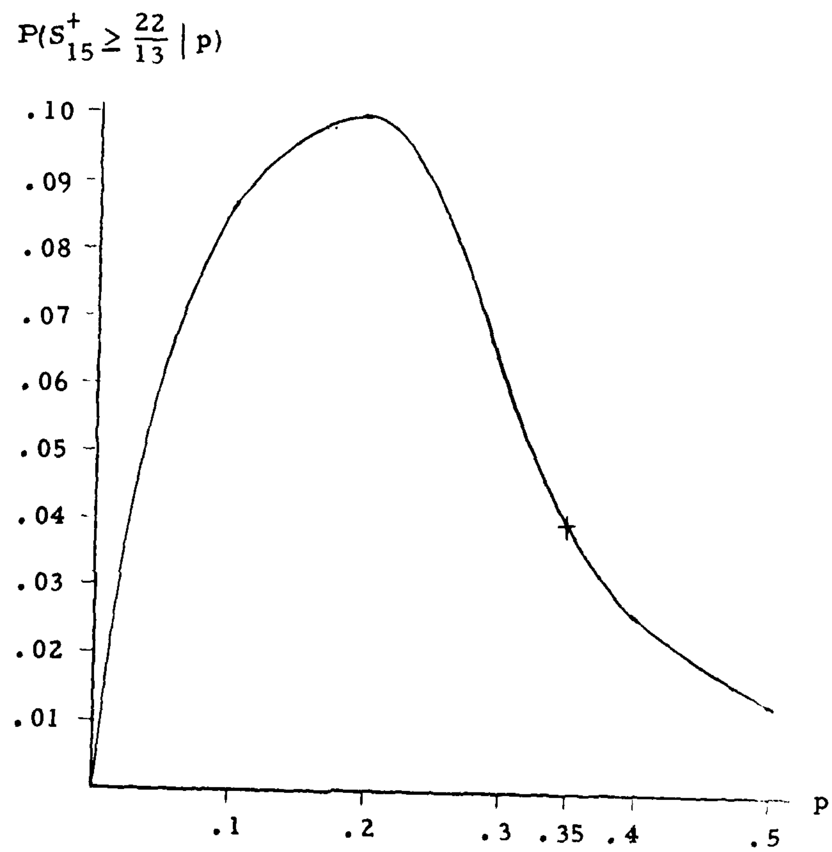


Figure 10. $P(S_{15}^{+} \geq \frac{22}{13} \mid p)$ *vs.* p .

In particular for $s^+ > 1$ maximums occur at $p = \frac{1}{2}$ for the following s^+ values:

$n = 5: \quad 1 < s^+ \leq 2$

$n = 6: \quad 1 < s^+ \leq \frac{8}{5} = 1.6$

$n = 7: \quad 1 < s^+ \leq \frac{8}{5} = 1.6$

$n = 8: \quad 1 < s^+ \leq \frac{10}{7} \doteq 1.429$

$n = 9: \quad 1 < s^+ \leq \frac{10}{7} \doteq 1.429$

$n = 10: \quad 1 < s^+ \leq \frac{4}{3} \doteq 1.333$

$n = 15: \quad 1 < s^+ \leq \frac{16}{13} \doteq 1.231$

$n = 20: \quad 1 < s^+ \leq \frac{22}{19} \doteq 1.158$

It appears as though the maximums are occurring at $p = \frac{1}{2}$ for smaller s^+ values as n increases. Also, as s^+ decreases from 2 the local maximums are monotonically approaching $p = \frac{1}{2}$ for each n. Once the maximum occurs at $p = \frac{1}{2}$ for an s^+ value, it occurs at $p = \frac{1}{2}$ for all remaining smaller values $s^+ > 1$.

For $s^+ \leq 1$ minimums occur at $p = \frac{1}{2}$ for the following s^+ values:

$n = 5: \quad s^+ \leq 1$

$n = 6: \quad s^+ \leq \frac{2}{3} \doteq .667$

$n = 7: \quad s^+ \leq \frac{5}{7} \doteq .714$

$n = 8: \quad s^+ \leq \frac{6}{8} \doteq .75$

$n = 9: \quad s^+ \leq \frac{2}{3} \doteq .667$

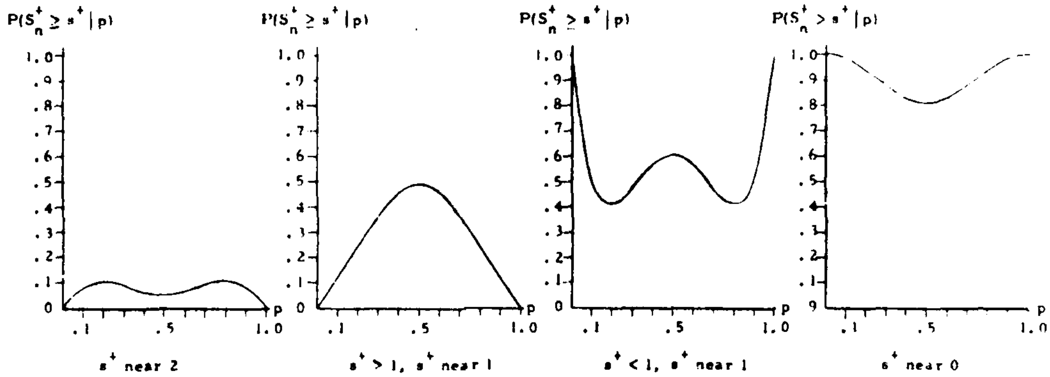
$n = 10: \quad s^+ \leq \frac{3}{5} \doteq .6$

$$n = 15: \quad s^+ \leq \frac{3}{5} = .6$$

$$n = 20: \quad s^+ \leq \frac{8}{14} \doteq .571$$

Once the minimum occurs at $p = \frac{1}{2}$ for an s^+ value, it occurs at $p = \frac{1}{2}$ for all remaining smaller values $s^+ < 1$.

Graphically, $P(S_n^+ \geq s^+ \mid p)$ vs. p is changing as s^+ decreases from 2 to 0 in the following way.



This behavior and Theorem 11 suggest that the conjecture following that theorem might be true.

The next two theorems investigate the behavior of the test statistic as $n \longrightarrow n+1$.

Theorem 12. $S_n^+(X_1, \cdots, X_n) \leq S_{n+1}^+(X_1, \cdots, X_n, -1)$ for all n.

Proof. Suppose $S_n^+(X_1, \cdots, X_n) = \frac{1}{j} \sum_{i=1}^{j} X_i - \frac{1}{n-j} \sum_{i=j+1}^{n} X_i$ where $j \neq 0, n$. Since $n+1-j > n-j$ and

$$\sum_{i=j+1}^{n+1} X_i = \sum_{i=j+1}^{n} X_i - 1 < \sum_{i=j}^{n} X_i \quad \text{we have}$$

$$\frac{1}{n+1-j} \sum_{i=j+1}^{n+1} X_i < \frac{1}{n-j} \sum_{i=j+1}^{n} X_i$$

$$\Longrightarrow \frac{1}{j} \sum_{i=1}^{j} X_i - \frac{1}{n-j} \sum_{i=j+1}^{n} X_i < \frac{1}{j} \sum_{i=1}^{j} X_i - \frac{1}{n+1-j} \sum_{i=j+1}^{n+1} X_i$$

$$\Longrightarrow S_n^+(X_1, \cdots, X_n) < S_{n+1}^+ (X_1, \cdots, X_n, -1) .$$

For those sequences $X_1, \cdots, X_n$ for which $S_n^+ (X_1, \cdots, X_n) = -\frac{1}{n} \sum_{i=1}^{n} X_i$ we have $-\frac{1}{n} \sum_{i=1}^{n} X_i < -\frac{1}{n+1} \sum_{i=1}^{n+1} X_i$. For those sequences $X_1, \cdots, X_n$ for which $S_n^+(X_1, \cdots, X_n) = \frac{1}{n} \sum_{i=1}^{n} X_i$ we have $S_{n+1}^+(X_1, \cdots, X_n, -1) \geq \frac{1}{n} \sum_{i=1}^{n} X_i - \frac{1}{1} (-1) = S_n^+(X_1, \cdots, X_n) + 1$. ■

<u>Theorem 13.</u> $S_n^+(X_1, \cdots, X_n) \geq S_{n+1}^+(X_1, \cdots, X_n, 1)$ except possibly when $S_{n+1}^+(X_1, \cdots, X_n, 1) = \frac{1}{n+1} \sum_{i=1}^{n+1} X_i < 1$.

<u>Proof.</u> Suppose $S_{n+1}^+(X_1, \cdots, X_n, 1) = \frac{1}{j} \sum_{i=1}^{j} X_i - \frac{1}{n+1-j} \sum_{i=j+1}^{n+1} X_i$ for some $j \neq 0, n, n+1$. Since $\sum_{i=j+1}^{n} X_i \leq n-j$ we have

$$\{(n+1-j) - (n-j)\} \sum_{i=j+1}^{n} X_i \leq n-j$$

$$\Longrightarrow (n+1-j) \sum_{i=j+1}^{n} X_i \leq (n-j) \sum_{i=j+1}^{n} X_i + (n-j) X_{n+1}$$

$$\Longrightarrow (n+1-j) \sum_{i=j+1}^{n} X_i \leq (n-j) \sum_{i=j+1}^{n+1} X_i$$

$$\Longrightarrow \frac{1}{j} \sum_{i=1}^{j} X_i - \frac{1}{n+1-j} \sum_{i=j+1}^{n+1} X_i \leq \frac{1}{j} \sum_{i=1}^{j} X_i - \frac{1}{n-j} \sum_{i=j+1}^{n} X_i$$

$$\Longrightarrow S^+_{n+1}(X_1, \cdots, X_n, 1) \leq S^+_n(X_1, \cdots, X_n)$$

for those sequences where $S^+_{n+1}(X_1, \cdots, X_n, 1)$ does not occur at $j = 0, n, n+1$.

Suppose S^+_{n+1} occurs at $j = 0$; then

$$-n \leq - \sum_{i=1}^{n} X_i$$

$$\Longrightarrow -n \sum_{i=1}^{n} X_i - n \leq -n \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} X_i$$

$$\Longrightarrow -n \left\{ \sum_{i=1}^{n} X_i + 1 \right\} \leq -(n+1) \sum_{i=1}^{n} X_i$$

$$\Longrightarrow -n \sum_{i=1}^{n+1} X_i \leq -(n+1) \sum_{i=1}^{n} X_i$$

$$\Longrightarrow -\frac{1}{n+1} \sum_{i=1}^{n+1} X_i \leq -\frac{1}{n} \sum_{i=1}^{n} X_i$$

$$\Rightarrow S^+_{n+1}(X_1, \cdots, X_n, 1) \leq S^+_n(X_1, \cdots, X_n) .$$

$S^+_{n+1}(X_1, \cdots, X_n, 1)$ cannot occur at $j = n$. See Theorem 10. ∎

Study of the computations for $P(S^+_n \geq s^+ \mid p)$ suggests that for all n and p and useful α levels it is usually true that $P(S^+_n \geq s^+ \mid p) \geq P(S^+_{n+1} \geq s^+ \mid p)$; when the inequality is not true the difference between the probabilities is quite small. Thus if $S^+_n(x_1, \cdots, x_n)$ is in the critical region for a size-α test then $S^+_{n+1}(x_1, \cdots, x_n, -1)$ is in the critical region for an approximate size-α test by Theorem 12. For the exceptional sequences in Theorem 13, $S^+_{n+1}(x_1, \cdots, x_n, 1) < 1$ and these sequences are not in the critical region for any reasonable size-α test for any n. For those sequences for which the inequality in Theorem 13 is true, we cannot say in general that $S^+_{n+1}(x_1, \cdots, x_n, 1)$ is not in the critical region for an approximate size-α test when $S^+_n(x_1, \quad , x_n)$ is not in the critical region for a size-α test. One must consider more closely the particular sequence $x_1, \cdots, x_n$ and the probability p.

More generally, computation indicates that for $n>20$ we can test H_0 vs. H^+_1 as follows. Suppose $S^+_n = s^+_n$; choose the largest value s^+ of S^+_{20} so that $s^+ \leq s^+_n$. If s^+ is in the critical region for a size-α test for sample size 20 then s^+_n will be in the critical region for an approximate size-α test for sample size n.

Theorem 15. If $S_n(X_1, \cdots, X_n) = \left| \frac{1}{j} \sum_{i=1}^{j} X_i - \frac{1}{n-j} \sum_{i=j+1}^{n} X_i \right|$ and $j \neq 0, n$, then $X_j = -X_{j+1}$.

Hence to evaluate S_n for a sample sequence one need only calculate $\left| \frac{1}{r} \sum_{i=1}^{r} x_i - \frac{1}{n-r} \sum_{i=r+1}^{n} x_i \right|$ for those r for which $x_r = -x_{r+1}$, and if each value calculated is less than 1 then also for $r = 0$.

For $n = 1$-$10, 15, 20$, $o(\Omega_{S_n})$ vs. n is plotted in Figure 11. As with $o(\Omega_{S_n^+})$, $o(\Omega_{S_n})$ increases rapidly with n. $\log_2(o(\Omega_{S_n}))$ vs. $\log_2 n$ is plotted in Figure 12. For $n = 4$-$10, 15, 20$ the graph is almost a straight line. The best least squares straight line fitting these points is $\log_2(o(\Omega_{S_n})) = 2.3221 \log_2 n - 3.0805$ so an approximation to $o(\Omega_{S_n})$ may be found by $o(\Omega_{S_n}) = 2^{-3.0805} n^{2.3221}$.

In Section 2.3 we found an approximation to $o(\Omega_{S_n^+})$ as $o(\Omega_{S_n^+}) = 2^{-2.456} n^{2.296}$ and since $o(\Omega_{S_n^+}) > o(\Omega_{S_n})$ for $n \geq 2$, one cannot approximate the orders of the sample spaces for $n > 20$ with confidence using these formulas.

We next investigate the behavior of the probability density S_n and its mode, mean, variance and median for varying n and p. In Figures 13.1, 13.2 are graphed $P(S_n = s \mid p)$ vs. s for $n = 10$ and 20, for $p = .1, .2, .3, .4, .5$ and for those s for which

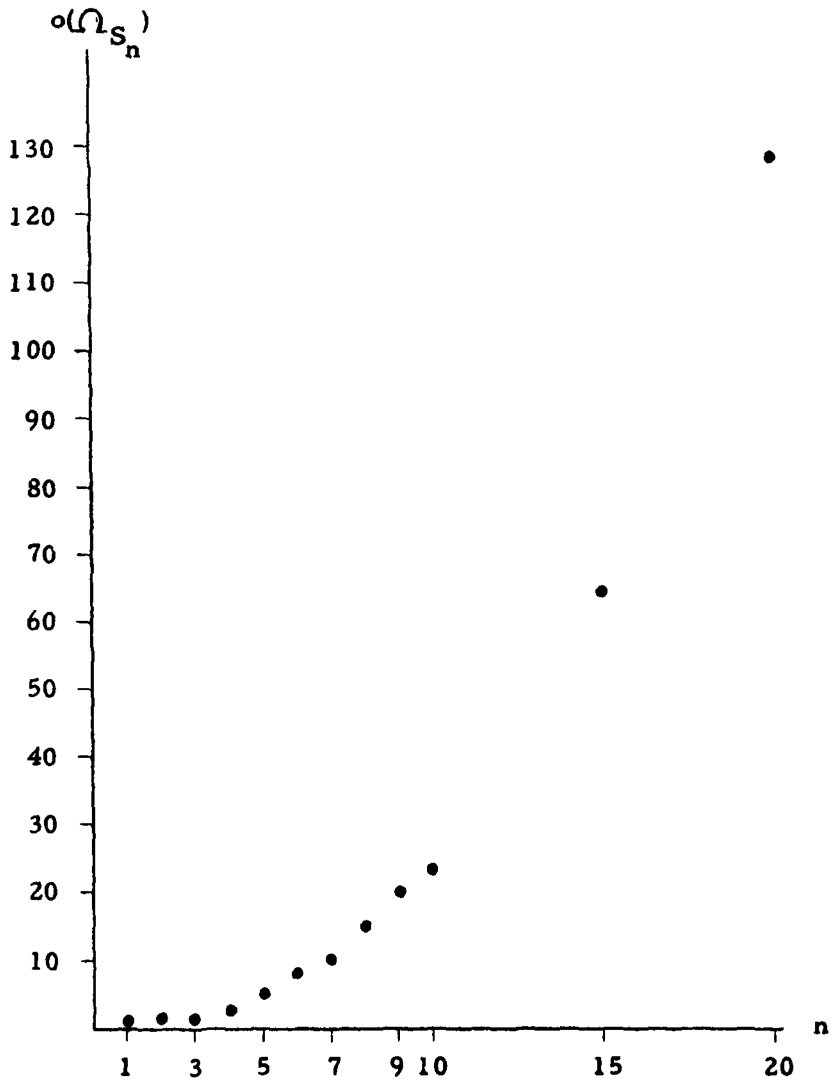


Figure 11. $o(\Omega_{S_n})$ vs. n .

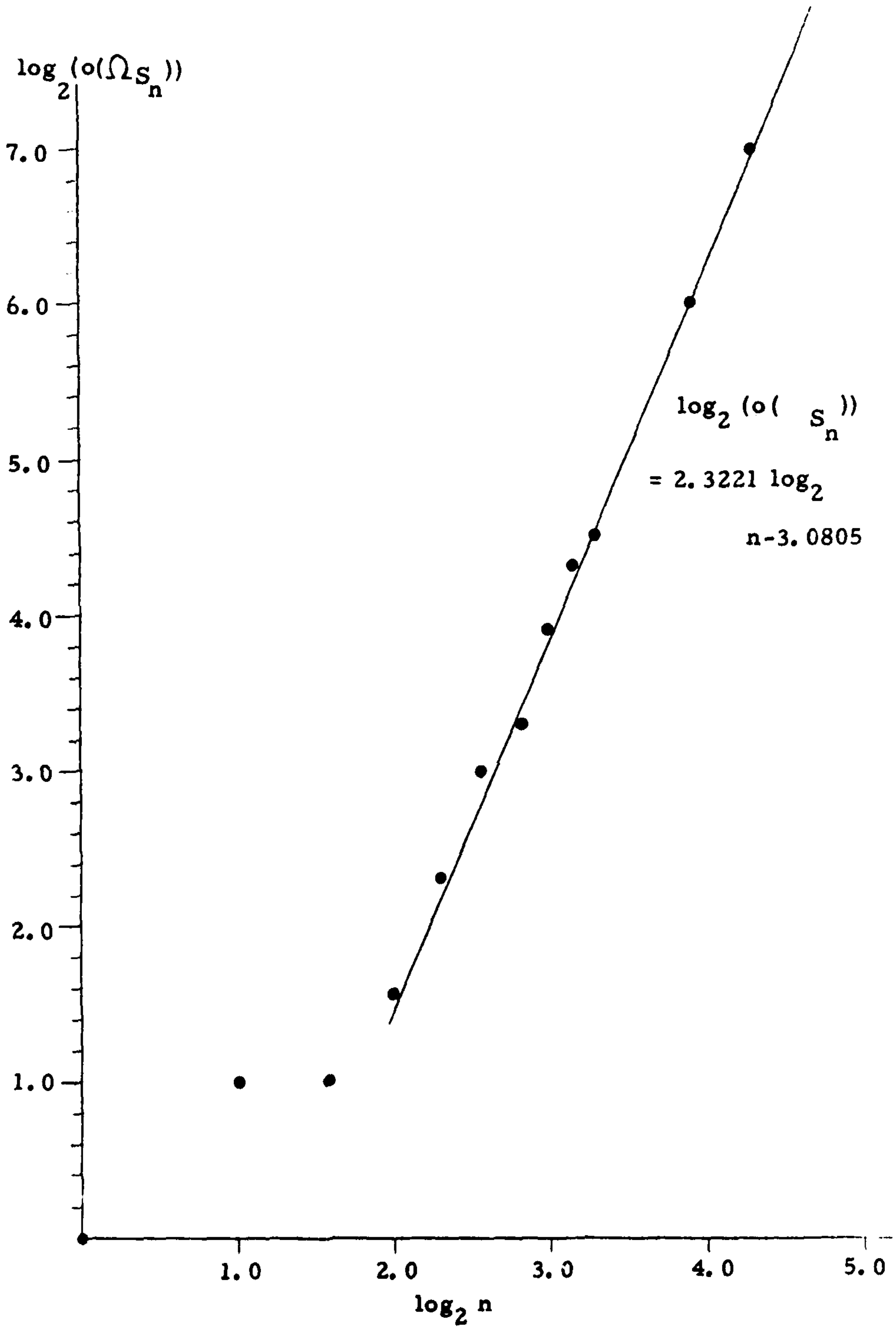


Figure 12. $\log_2(o(\Omega_{S_n}))$ vs. $\log_2 n$.

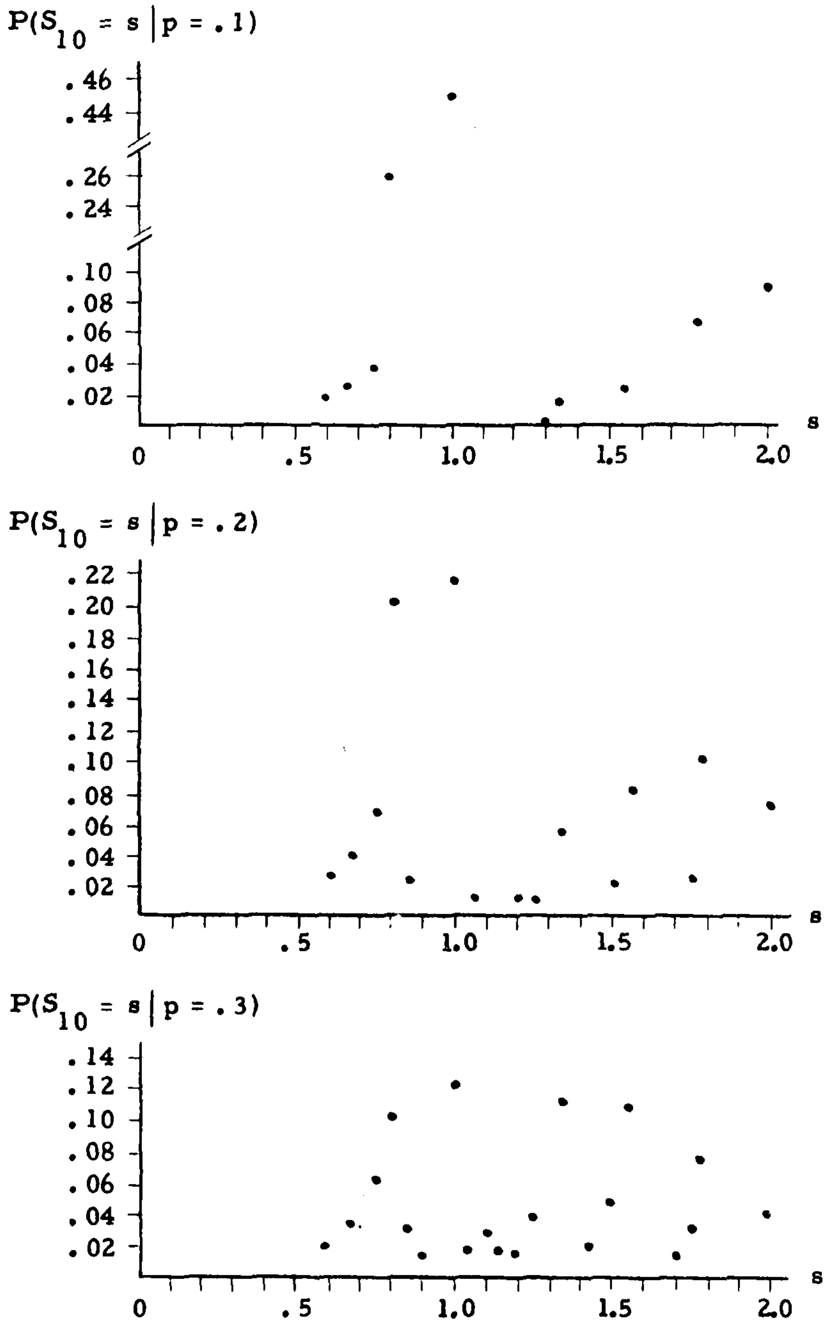
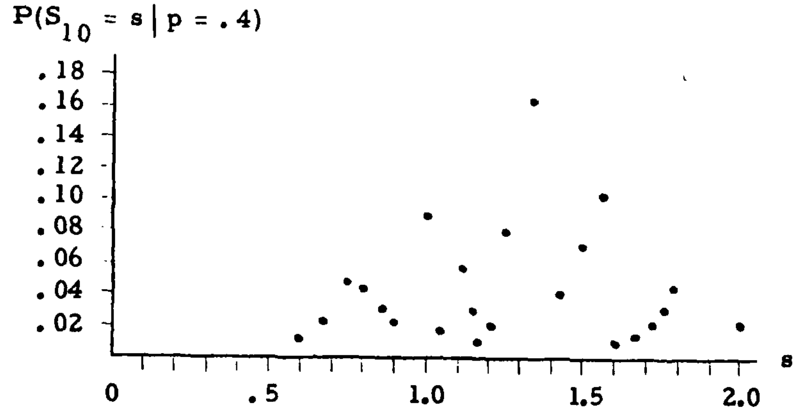


Figure 13.1 $P(S_{10} = s \mid p)$ vs. s.



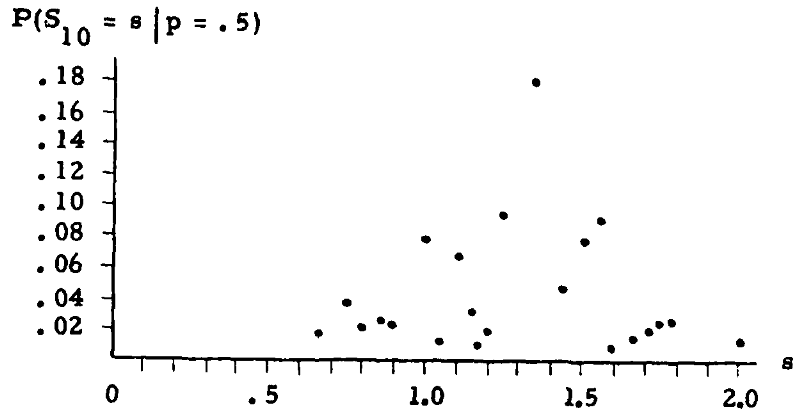


Figure 13.1. (Continued).

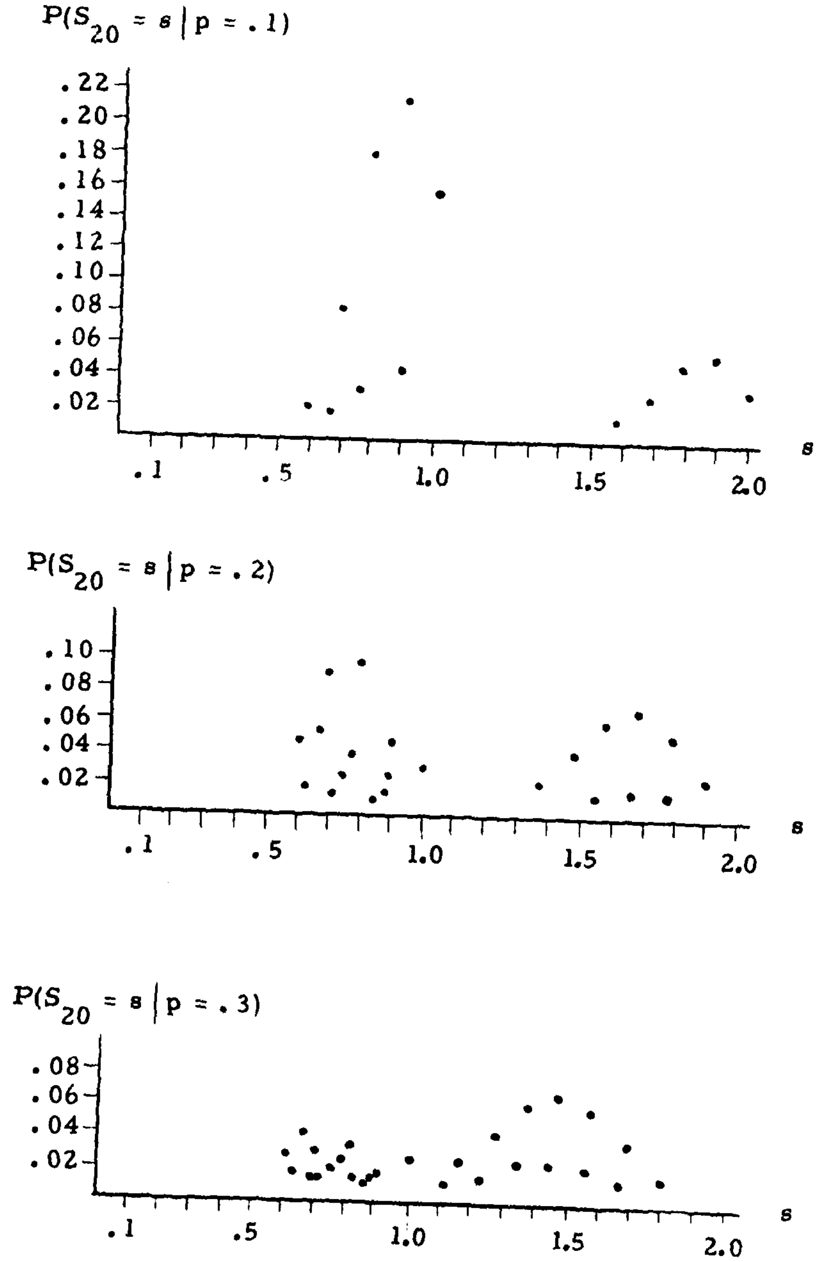
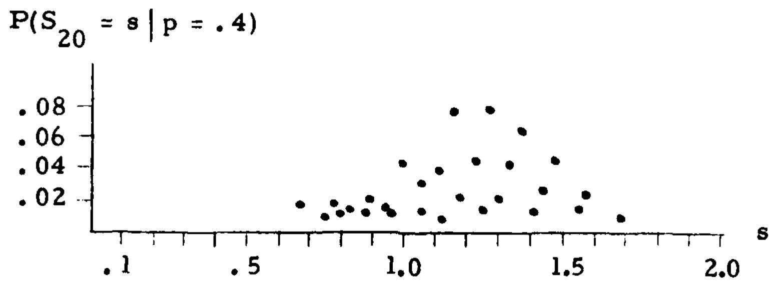


Figure 13.2 $P(S_{20} = s \mid p)$ vs. s .



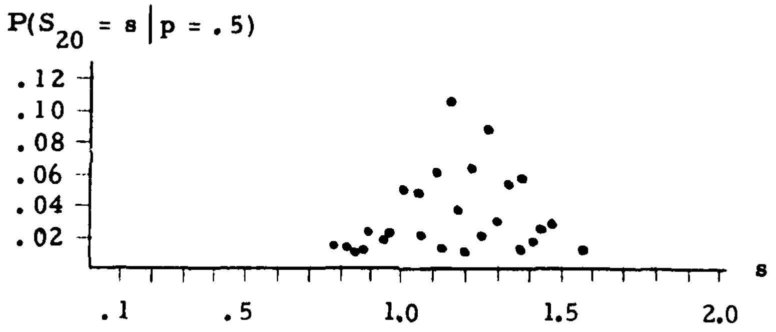


Figure 13. 2 (Continued)

$P(S_n = s \mid p) \geq .01$. By Theorem 6 the probability densities for $p = .6, .7, .8, .9$ are the same as those for $1-p$ and hence are omitted.

For each n plotted as $p: .1 \longrightarrow .5$ the number of "large" probability masses are shifting away from extreme s values. The same phenomenon occurs for S_n^+ for large extreme values.

The mode of S_n behaves as the mode of S_n^+ in regard to the shifting through s values and the change in the mode's probability as p varies.

Figure 14 illustrates the behavior of $E(S_n \mid p)$ for $n = 5\text{-}10, 15, 20$ as p varies. The mean is symmetrical about $p = \frac{1}{2}$ for all n due to Theorem 6 and $E(S_n \mid p=0) = 1$. For each n plotted $E(S_n \mid p)$ increases as $p: 0 \longrightarrow \frac{1}{2}$ and for each p, $E(S_n \mid p)$ is a decreasing function of n . Hence the range of expected values over p for given n decreases as n increases; for large n and p near $\frac{1}{2}$ the expected values are fairly constant.

In comparing $E(S_n \mid p)$ to $E(S_n^+ \mid p)$ for $n = 5\text{-}10, 15, 20$ note that both have at least local maximums at $p = \frac{1}{2}$ and both in general are decreasing functions in n for fixed p . Also note that $E(S_n \mid p) \geq E(S_n^+ \mid p)$ for all p and n .

$\text{Var}(S_n \mid p)$ vs. p is sketched in Figure 15 for $n = 5\text{-}10, 15, 20$; $\text{Var}(S_n \mid p) = \text{Var}(S_n \mid 1-p)\ \forall p, n$ and $\text{Var}(S_n \mid p=0) = 0\ \forall n$. For each n plotted local maximums occur at approximately $p = .2$

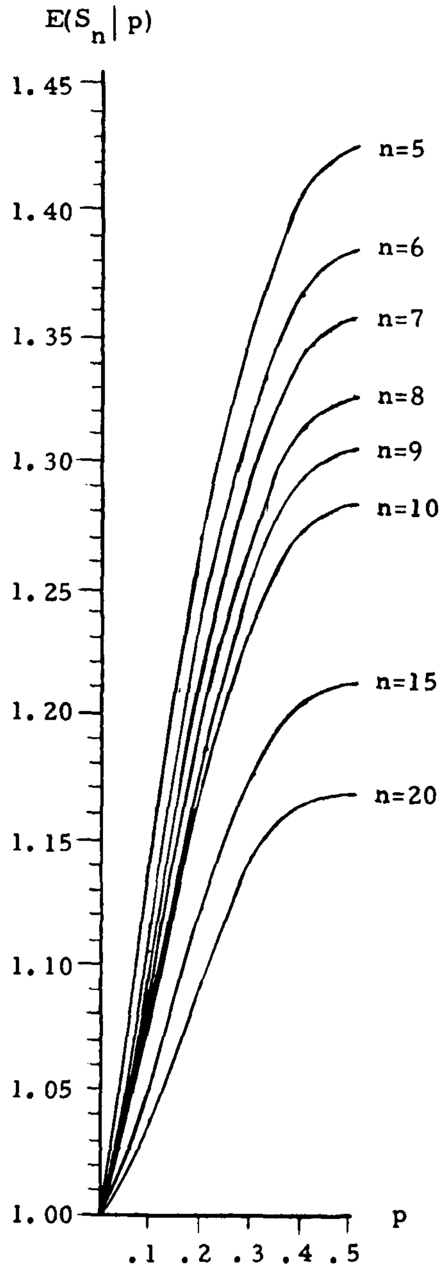


Figure 14. $E(S_n \mid p)$ vs. p .

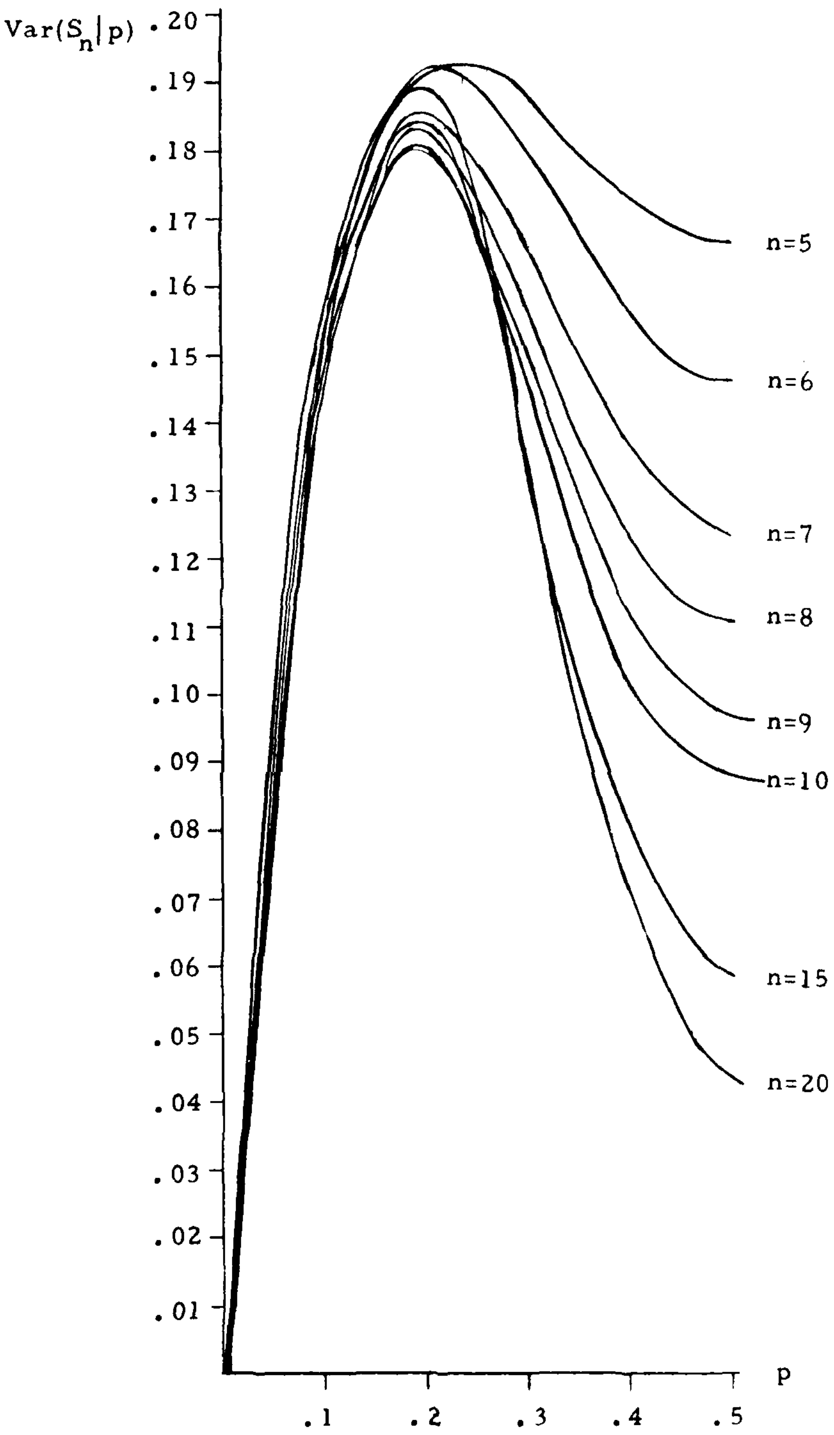


Figure 15. Var (S_n|p) vs. p .

AD-A087 111 COLORADO STATE UNIV FORT COLLINS DEPT OF STATISTICS F/G 12/2
TESTING FOR A CHANGE POINT.(U)
APR 80 N J FARDEN N00014-75-C-0518
UNCLASSIFIED TR-19 NL

2 OF 2
AD A
087111

END
DATE
FILMED
9-80
DTIC

and $.8$ and a local minimum occurs at $p = \frac{1}{2}$; such a behavior is expected from the shift in probabilities for S_n as $p: 0 \rightarrow \frac{1}{2}$. The variance depends heavily on p and for large n is quite small for p near $\frac{1}{2}$. For all n the variance is approximately the same for fixed p, where $p \leq .2$ and $p \geq .8$.

In Figure 16 is graphed $\text{med}(S_n \mid p)$ vs. p for $n = 5, 10, 15, 20$. The median is symmetric in p and $1-p$; $\text{med}(S_n \mid p=0) = 1$ for all n and maximums occur at $p = \frac{1}{2}$ for each n plotted; $E(S_n \mid p)$ behaves similarly. For larger n however the median decreases as p increases from 0 and then increases as p nears $\frac{1}{2}$; the mode exhibited this behavior.

Since $\text{med}(S_n \mid p) \geq \text{med}(S_n^+ \mid p)$ for all n and p, $\text{med}(S_n \mid p = \frac{1}{2}) \geq 1$ for all n by Theorem 11. It is tempting to conjecture that $\text{med}(S_n \mid p = \frac{1}{2}) \geq \text{med}(S_n \mid p)$ for all p and n.

Critical regions for the test H_0 vs. H_1 are contained in Table 17 for $n = 5, 10, 15, 20$ and $p = .1, .2, .3, .4, .5$; critical regions for $1-p$ are available from the table by Theorem 6. As with the test H_0 vs. H_1^+ interpolation can be used to find critical regions where p assumes a value other than those listed in the tables. Again interpolation for a significance test between two s values in the table is hazardous.

For fixed n and s, $P(S_n \geq s \mid p)$ as a function of p behaves as did $P(S_n^+ \geq s^+ \mid p)$. Properties (A) and (B) on page 70 are

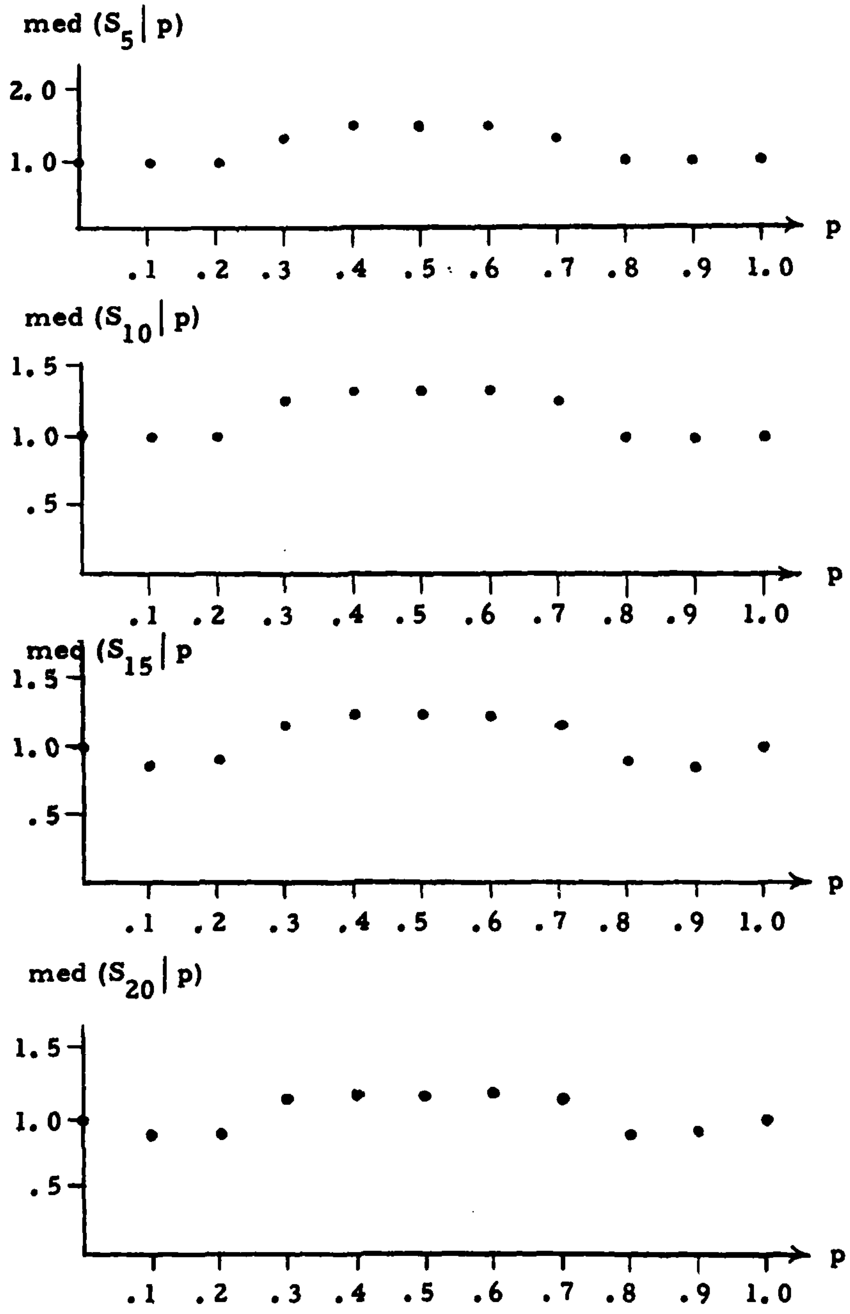


Figure 16. med $(S_n|p)$ vs. p .

Table 17. $P(S_n \geq s \mid p)$.

n=5

s \ p	.1	.2	.3	.4	.5
2/1	.148	.218	.244	.250	.250
4/3	.204	.397	.552	.653	.688

n=6

s \ p	.1	.2	.3	.4	.5
2/1	.133	.175	.174	.162	.156
8/5	.179	.296	.353	.372	.375
6/5	.206	.405	.571	.680	.719

n=7

s \ p	.1	.2	.3	.4	.5
2/1	.120	.140	.123	.102	.094
5/3	.173	.260	.270	.247	.234
6/5	.206	.409	.580	.694	.734

n=8

s \ p	.1	.2	.3	.4	.5
2/1	.108	.112	.086	.063	.055
12/7	.166	.228	.207	.162	.141
3/2	.175	.276	.309	.309	.305
8/7	.206	.410	.586	.707	.750

n=9

s \ p	.1	.2	.3	.4	.5
2/1	.097	.089	.060	.039	.031
14/8	.159	.199	.158	.105	.082
12/7	.165	.224	.199	.151	.129
8/5	.166	.235	.230	.204	.191
11/10	.205	.409	.587	.713	.758

Table 17 (Continued).

n=10

s \ p	.1	.2	.3	.4	.5
2/1	.087	.072	.042	.024	.018
16/9	.152	.172	.121	.067	.047
14/8	.158	.196	.153	.097	.074
12/7	.159	.201	.166	.118	.098
8/5	.160	.206	.180	.143	.127
14/9	.182	.287	.291	.246	.221
22/21	.208	.419	.605	.735	.781

n=15

s \ p	.1	.2	.3	.4	.5
2/1	.051	.023	.007	.002	.001
26/14	.115	.078	.029	.007	.002
24/13	.122	.092	.038	.011	.004
14/8	.123	.096	.044	.016	.009
12/7	.163	.174	.098	.037	.019
22/13	.167	.192	.119	.050	.027
5/3	.167	.196	.127	.058	.034
16/10	.168	.198	.132	.067	.044
11/7	.183	.266	.211	.115	.074
14/9	.183	.266	.212	.117	.077
20/13	.185	.280	.241	.147	.103
41/28	.185	.285	.254	.170	.128
13/9	.185	.286	.260	.182	.144
20/14	.189	.326	.341	.263	.215
14/13	.205	.399	.576	.715	.770

Table 17 (Continued).

n=20

s \ p	.1	.2	.3	.4	.5
2/1	.030	.008	.001	.000	.000
36/19	.083	.033	.006	.001	.000
34/18	.089	.039	.008	.001	.000
9/5	.089	.041	.010	.002	.001
34/19	.137	.093	.028	.005	.001
16/9	.142	.105	.035	.007	.002
22/13	.143	.108	.040	.010	.003
32/19	.170	.175	.080	.020	.006
5/3	.173	.190	.095	.027	.009
28/17	.173	.193	.101	.031	.011
38/24	.173	.194	.104	.035	.015
30/19	.184	.254	.165	.061	.025
142/91	.184	.254	.165	.062	.026
20/13	.185	.267	.189	.078	.036
3/2	.186	.271	.199	.092	.049
28/19	.189	.311	.270	.139	.076
16/11	.189	.311	.271	.142	.080
26/18	.189	.320	.296	.169	.103
140/99	.189	.320	.297	.172	.107
24/17	.189	.321	.305	.187	.123
136/99	.190	.323	.310	.199	.139
26/19	.191	.344	.372	.266	.194
22/21	.206	.396	.563	.709	.773

satisfied by S_n. For s near 2 (A1) is satisfied, as s approaches 1 (A2) is satisfied, and if s_1 is the largest value for which (A2) is satisfied then for each s, $s_1 \leq s < 1$, (A2) is satisfied. For $s \leq 1$ either (B1) or (B2) is satisfied.

For very large s we have that $P(S_n \geq s \mid p) \doteq 2P(S_n^+ \geq s \mid p)$. One expects near equality for large s since if $X_1, \cdots, X_n$ has $S_n = S_n^+ = s$ then $X_n, \cdots, X_1$ also has $S_n = s$. The equality breaks down when $X_1, \cdots, X_n = X_n, \cdots, X_1$.

Study of computations suggests that for all n and p and useful α levels it is usually true that $P(S_n \geq s \mid p) \geq P(S_{n+1} \geq s \mid p)$; when the inequality is not true the difference between probabilities is quite small. Tests for $n > 20$ can be obtained by using this fact.

2.5 The Covariance of S_n^+ and R_n^+

Recall that in general for each value $s^+ \epsilon \Omega_{S_n^+}$ there are only two points j, $n-j \epsilon \Omega_{R_n^+}$ at which the value is assumed; exceptions occur for $s^+ = 2$ and a few more "popular" values such as $s^+ = 1, 0$. A measure for further investigating the relationship between S_n^+ and R_n^+ is provided by their covariance. Since ρ is estimated only when the null hypothesis is rejected, the covariance of S_n^+ and R_n^+ given that $S_n^+ \geq s^+$ is of particular interest. Theorem 17 describes the behavior of the conditional covariance; Lemma 16 will be useful.

<u>Lemma 16.</u> For each n, $p \in (0, 1)$ and s^+, $E(R_n^+ \mid p, S_n^+ \geq s^+) + E(R_n^+ \mid 1-p, S_n^+ \geq s^+) = n$; for each $p \in \left[0, \frac{1}{2}\right]$ and $s^+ > 1$, $E(R_n^+ \mid p, S_n^+ \geq s^+) \leq E(R_n^+ \mid 1-p, S_n^+ \geq s^+)$; for each n and s^+, $E(R_n^+ \mid p = 0, S_n^+ \geq s^+) = 0$.

<u>Proof.</u> Note that

$$E(R_n^+ \mid p, S_n^+ \geq s^+) + E(R_n^+ \mid 1-p, S_n^+ \geq s^+)$$

$$= \sum_{j=0}^{n} j \{ P(R_n^+ = j \mid p, S_n^+ \geq s^+) + P(R_n^+ = j \mid 1-p, S_n^+ \geq s^+) \}$$

$$= \sum_{j=0}^{n} j \{ P(R_n^+ = j \mid p, S_n^+ \geq s^+) + P(R_n^+ = n-j \mid p, S_n^+ \geq s^+) \}$$

$$= \sum_{j=0}^{n} \{ j \, P(R_n^+ = j \mid p, S_n^+ \geq s^+) + (n-j) \, P(R_n^+ = j \mid p, S_n^+ \geq s^+) \}$$

$$= n \sum_{j=0}^{n} P(R_n^+ = j \mid p, S_n^+ \geq s^+) = n .$$

Let $d_k^+(j \mid s^+)$ denote the number of sequences with k 1's, $S_n^+ \geq s^+$ and $R_n^+ = j$. It can be shown algebraically that $\sum_{j=0}^{i} d_k^+(j \mid s^+) \geq \sum_{j=0}^{i} d_k^+(n-j \mid s^+)$ for each $s^+ > 1$ and for each i and k where $0 \leq i \leq \left[\frac{n}{2}\right]$ and $0 \leq k \leq \left[\frac{n}{2}\right]$. Hence

$$\sum_{j=0}^{\left[\frac{n}{2}\right]} \sum_{k=0}^{\left[\frac{n}{2}\right]} (n-2j) \left[d_k^+(j \mid s^+) - d_k^+(n-j \mid s^+) \right] \left[p^k (1-p)^{n-k} - (1-p)^k p^{n-k} \right]$$

$$\geq 0$$

for $p \epsilon (0, \frac{1}{2})$ and $s^+ > 1$; this implies that

$$\sum_{j=0}^{n} j\, P(R_n^+ = j \mid 1-p,\ S_n^+ \geq s^+) \geq \sum_{j=0}^{n} j\, P(R_n^+ = j \mid p,\ S_n^+ \geq s^+) . \quad \blacksquare$$

It is quite likely true that $E(R_n^+ \mid p,\ S_n^+ \geq s^+) \leq E(R_n^+ \mid 1-p,$ $S_n^+ \geq s^+)$ for all s^+, n and $p \epsilon (0, \frac{1}{2})$; the proof would involve more complicated algebraic manipulations than for the situation where $s^+ > 1$.

Theorem 17. For each n, p and s^+, $\mathrm{Cov}\,(S_n^+,\ R_n^+ \mid p,\ S_n^+ \geq s^+)$ + $\mathrm{Cov}\,(S_n^+,\ R_n^+ \mid 1-p,\ S_n^+ \geq s^+) = 0$; for each n and p , $\mathrm{Cov}\,(S_n^+,\ R_n^+ \mid p,$ $S_n^+ = 2) = 0$ and for each n and s^+, $\mathrm{Cov}\,(S_n^+,\ R_n^+ \mid p = 0,\ S_n^+ \geq s^+) = 0$.

Proof. Observe that

$$\begin{aligned} E(S_n^+,\ R_n^+ \mid p,\ S_n^+ \geq s^+) &= \sum_{s^+, j} s^+ j\, P((S_n^+, R_n^+) = (s^+, j) \mid p, S_n^+ \geq s^+) \\ &= \sum_{s^+, j} s^+ j\, P((S_n^+, R_n^+) = (s^+, j) \mid 1-p, S_n^+ \geq s^+) \\ &= \sum_{s^+, j} \{n s^+ P((S_n^+, R_n^+) = (s^+, j) \mid 1-p, S_n^+ \geq s^+) \\ &\quad - (n-j)\, s^+\, P((S_n^+, R_n^+) = (s^+, j) \mid 1-p, S_n^+ \geq s^+)\} \\ &= n\, E(S_n^+ \mid 1-p, S_n^+ \geq s^+) - E(S_n^+ R_n^+ \mid 1-p, \\ &\qquad S_n^+ \geq s^+) . \end{aligned}$$

Also observe that $E(S_n^+ \mid p,\ S_n^+ \geq s^+) = E(S_n^+ \mid 1-p,\ S_n^+ \geq s^+)$ for all n, p and s^+. Apply Lemma 16. ■

For $s^+ = 2$ the conditional covariance of S_n^+ and R_n^+ is zero; for $s^+ < 2$ computation indicates that the conditional correlation may be positive or negative for $p\epsilon(0, \frac{1}{2})$. More specifically for $p\epsilon(0, \frac{1}{2})$ the correlation is negative for s^+, $1 < s^+ < 2$, and is positive when s^+ is quite small. The magnitude of the correlation of S_n^+ and R_n^+ for given p and s^+ is not large. Verification of this behavior of the correlation for arbitrary n and given p and s^+ appears to be intractable.

2.6 The Covariance of S_n and R_n

The conditional covariance of S_n and R_n given that $S_n \geq s$ is identically equal to zero for all n, p and s; S_n and R_n are definitely not independent random variables however. Lemma 18 indicates that the conditional mean of R_n does not depend on p or s.

Lemma 18. For each n, $p\epsilon(0, 1)$ and s, $E(R_n \mid p,\ S_n \geq s) = \frac{n}{2}$.

Proof.
$$E(R_n \mid p,\ S_n \geq s) = \sum_{j=0}^{n} j\, P(R_n = j \mid p,\ S_n \geq s)$$
$$= \frac{1}{2} \sum_{j=0}^{n} \{ j\, P(R_n = j \mid p,\ S_n \geq s) + (n-j)\, P(R_n = n-j \mid p,\ S_n \geq s)\}$$

$$= \frac{1}{2} \sum_{j=0}^{n} \{ j\, P(R_n = j \mid p,\ S_n \geq s)$$

$$+ (n-j)\, P(R_n = j \mid p,\ S_n \geq s) \}$$

$$= \frac{n}{2} \ . \qquad \blacksquare$$

<u>Theorem 19.</u> For each n, p and s, $\mathrm{Cov}\,(S_n, R_n \mid p, S_n \geq s) = 0$.

<u>Proof.</u> Note that

$$E(S_n R_n \mid p, S_n \geq s) = \sum_{s,j} sj\, P((S_n, R_n) = (s,j) \mid p, S_n \geq s)$$

$$= \frac{n}{2} \sum_{s,j} s\, P((S_n, R_n) = (s,j) \mid p, S_n \geq s)$$

$$= \frac{n}{2} \sum_{s} s\, P(S_n = s \mid p, S_n \geq s)$$

$$= \frac{n}{2}\, E(S_n \mid p, S_n \geq s) \ .$$

Apply Lemma 18. $\blacksquare$

2.7 Properties of R_n^+

Study of the change point random variable R_n^+ provides properties for the estimator $\hat{\rho}$ of the change point. Although R_n^+ is of interest in its own right, note that ρ is estimated only when H_0 is rejected as a study of the random variable $\hat{\rho}^+_{n,\alpha} = (R_n^+ \mid S_n^+ \geq s^+_{\alpha,p})$ is of primary concern. In this section properties of both random variables are investigated. First we examine the behavior of the

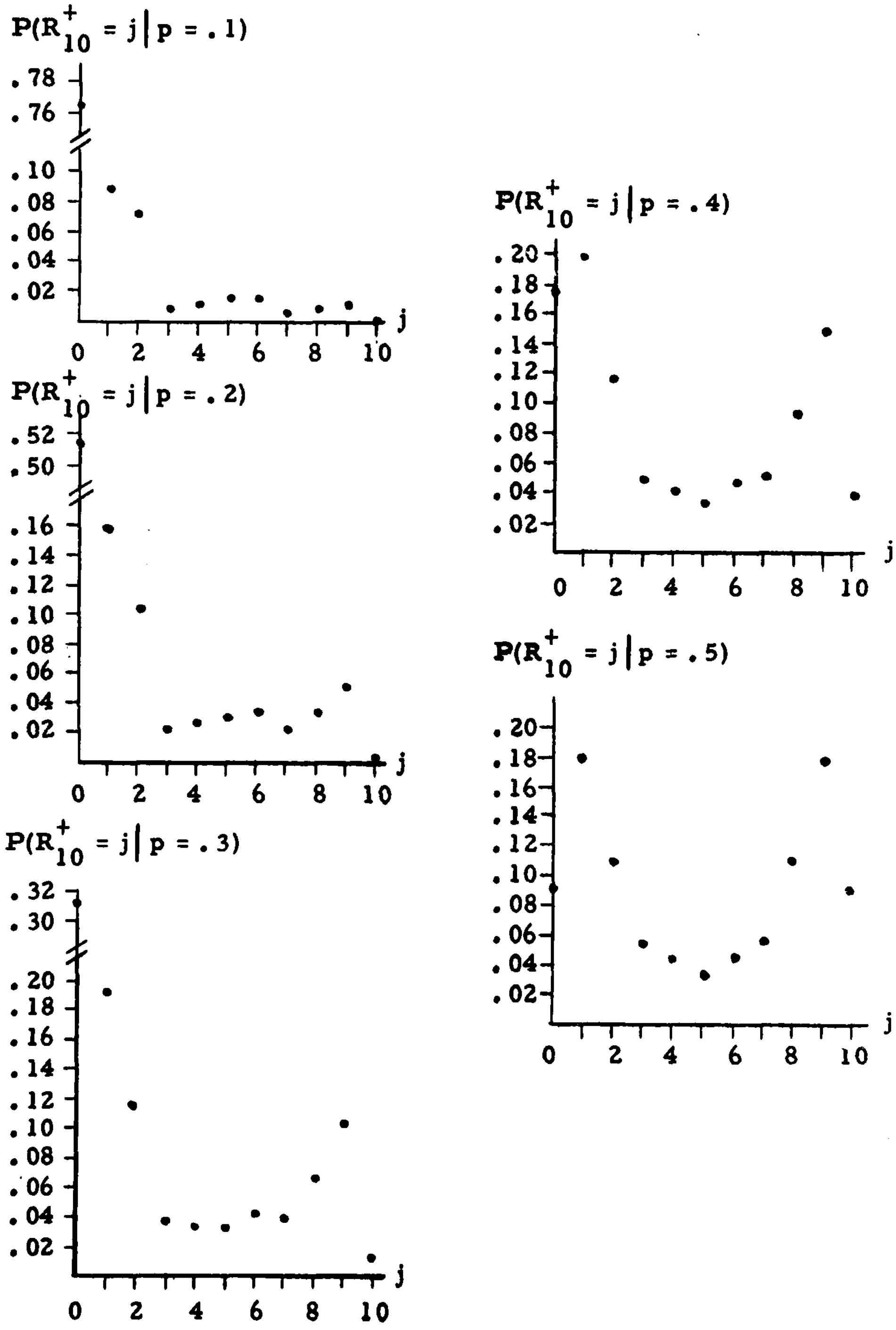


Figure 18.1 $P(R^{+}_{10} = j \mid p)$ vs. j .

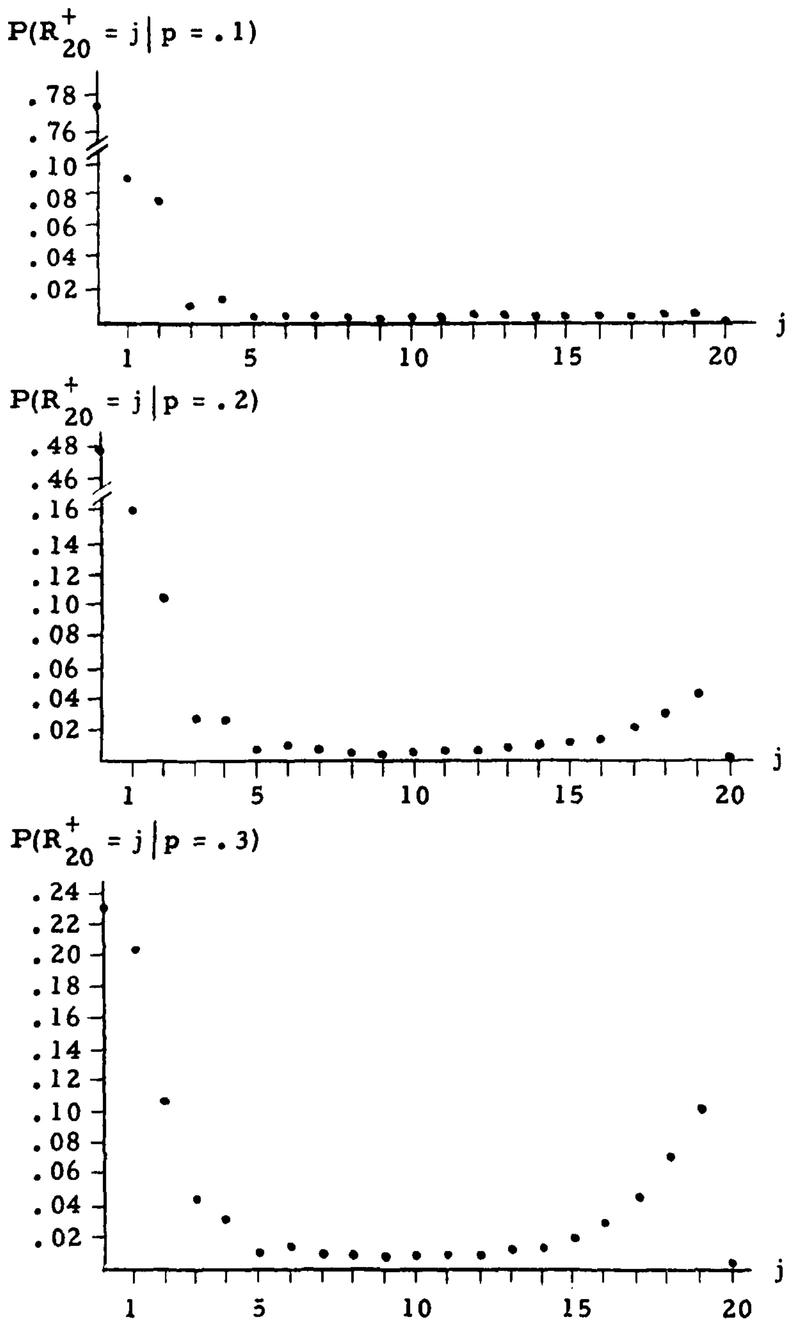
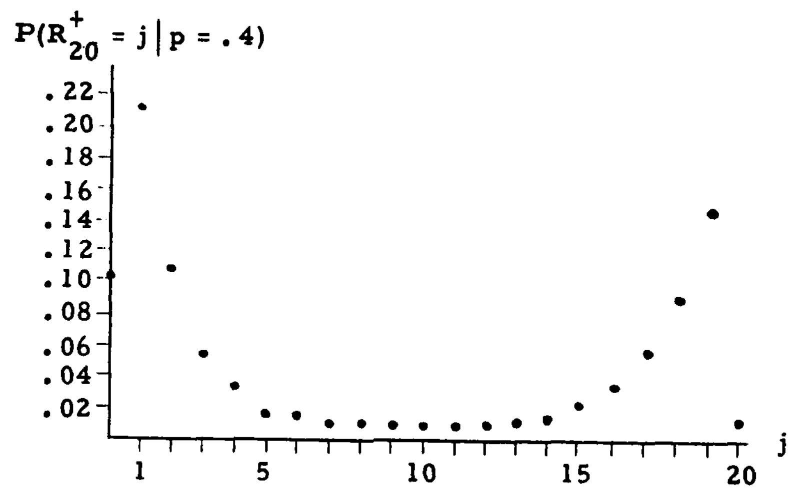


Figure 18.2 $P(R_{20}^{+} = j \mid p)$ vs. j.



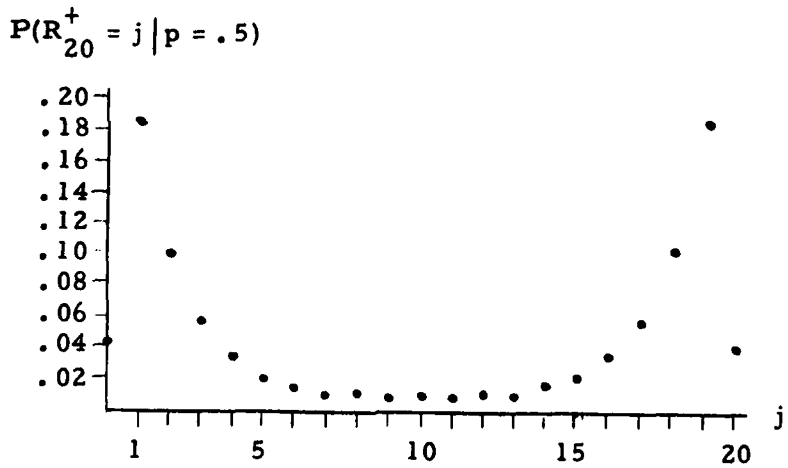


Figure 18.2 (Continued)

density of R_n^+ and its mean and variance as functions of n and p. Next $\hat{\rho}_{n,\alpha}^+$ is investigated and a comparison is made between $\rho_{n,\alpha}^+$ and the corresponding change point random variables for the generalized likelihood ratio test statistic (1.4.1) and Page's test statistic (1.4.2). The section is concluded proposing an alternative to randomizing for estimating ρ when H_0 is rejected and when J_n^+ is not uniquely defined by considering what happens to J_n^+ as $n \longrightarrow n+1$.

The behavior of the densities of R_n^+ is illustrated in Figures 18.1, 18.2 for $n = 10, 20$ and $p = .1, .2, .3, .4, .5$. The densities for $p = .6, .7, .8, .9$ are related to those for $1-p$ by $P(R_n^+ = j \mid p) = P(R_n^+ = n-j \mid 1-p)$ according to Theorem 2. We examine the densities separately for $R_n^+ = 0$ and $R_n^+ \epsilon \{1, \cdots, \left[\frac{n}{2}\right]\}$.

The sequences for which $R_n^+ = 0$ are of the type $T_n^k(0,0)$ where $k \leq \left[\frac{n}{2}\right]$ so $P(R_n^+ = 0 \mid p) = 1, 0$ when $p = 0, 1$ respectively. For sequences of this type it is easy to show that for each n, $P(R_n^+ = 0 \mid p)$ decreases as p increases at least for $p \epsilon \left[\frac{1}{2}, 1\right]$; it appears that this is also the tendency for $p \epsilon (0, \frac{1}{2})$. For each p, $P(R_n^+ = 0 \mid p)$ tends to decrease as n increases. The proportion of sequences for which $R_n^+ = 0$ is decreasing with n; note that .091 of the 2^{10} sequences have $R_{10}^+ = 0$ and about .042 of the 2^{20} sequence have $R_{20}^+ = 0$.

For $j \epsilon \{1, \cdots, \left[\frac{n-2}{2}\right]\}$ there are in general more sequences for which $R_n^+ = j$ then $R_n^+ = j+1$. To gain insight as to why this is the case consider those sequences for which $R_n^+ = 1$ and $R_n^+ = 2$. If

$R_n^+ = 1$ then $x_1 = 1$, $x_2 = -1$ and there may be k 1's in the last $n-2$ places, $0 \leq k \leq \left[\frac{(3n-2)(n-2)}{2n}\right]$; if $R_n^+ = 2$ then $x_1 = 1$ or -1, $x_2 = 1$, $x_3 = -1$ and there may be k 1's in the last $n-3$ places where $0 \leq k \leq \left[\frac{(3n-4)(n-2)}{4(n-1)}\right]$ if $x_1 = 1$ and $0 \leq k \leq \left[\frac{n-2}{2}\right]$ if $x_1 = -1$ (see Theorems 3 and 10). Thus the total number of sequences for which S_n^+ may occur at $j = 1$ is larger than for $j = 2$. Not all of the sequences described actually have $R_n^+ = 1$ or 2 however. For given k if many of the 1's are near the beginning of the sequence S_n^+ may occur at some larger point j ; note that the sequences for which $x_1 = -1$, $x_2 = 1$, $x_3 = -1$ are more sensitive in this situation and are more likely to have S_n^+ occurring $j > 2$.

Observe next that for those sequences for which $R_n^+ = j$ there generally are a larger proportion of 1's than -1's in $x_1, \cdots, x_j$ and a larger proportion of -1's than 1's in $x_{j+1}, \cdots, x_n$ since S_n^+ cannot be negative. Hence for p near 0 those points near $j = 1$ are most probable, for p near 1 those points near $j = n-1$ are most probable and for p near $\frac{1}{2}$ the extreme points j encompass most of the probability mass.

Thus for $j \epsilon \{1, \cdots, n-1\}$ and for given p the shape of the probability densities will be similar for abitrary n. Moreover calculation indicates that for the more probable extreme points j, $P(R_{n_1}^+ = j \mid p) = P(R_{n_2}^+ = j \mid p)$ for all n_1, n_2 and p . Among the less probable points j the remaining probability mass is similarly distributed for

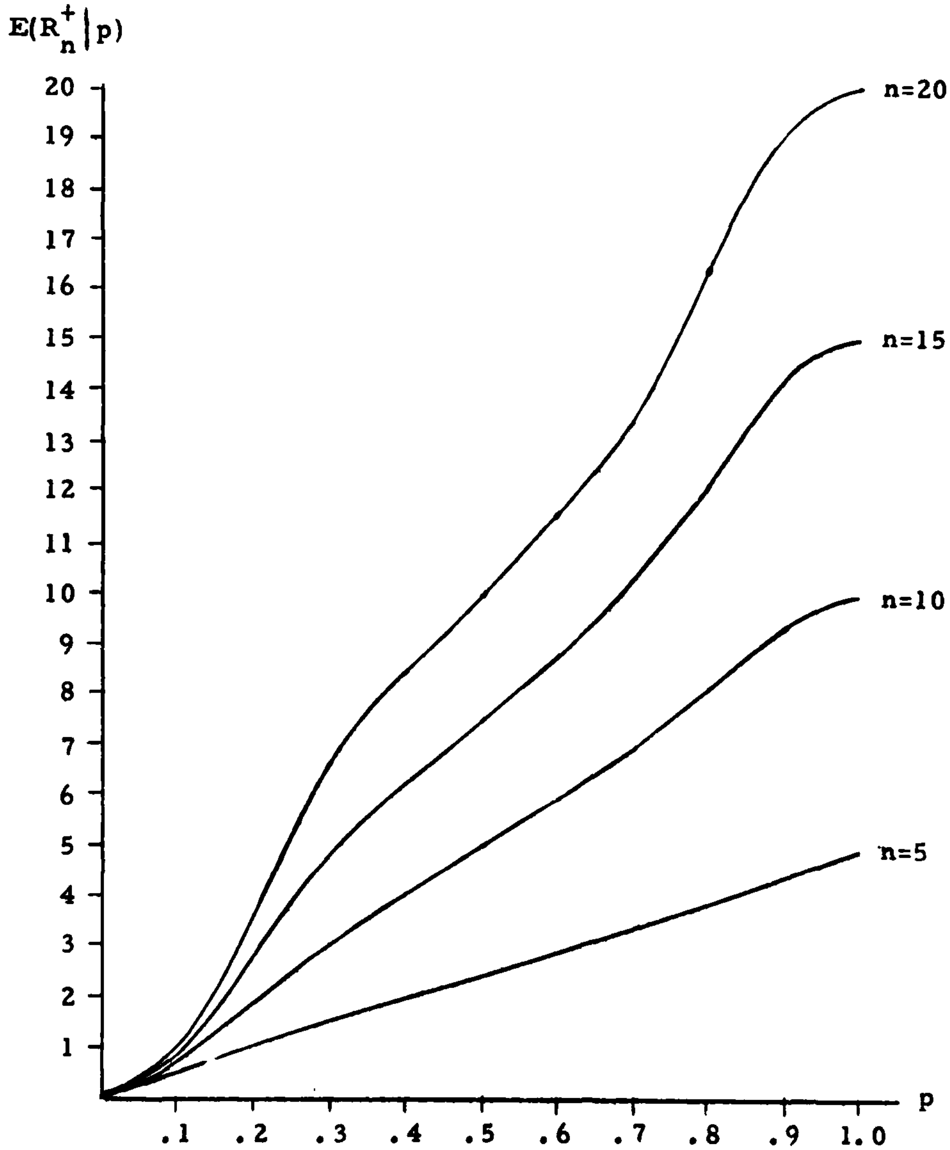


Figure 19. $E(R_n^+ | p)$ vs. p .

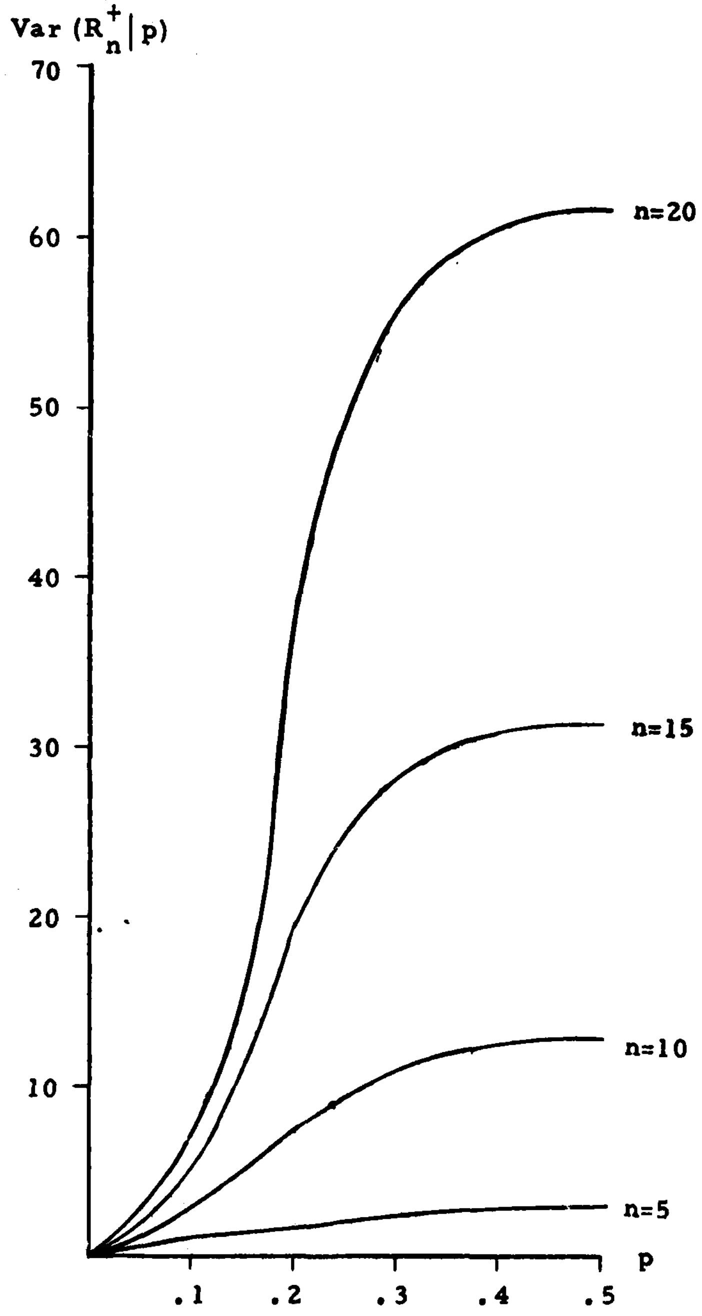


Figure 20. Var $(R_n^+|p)$ vs. p .

each n and p ; the probability of each of these points decreases as n increases for each p .

Figure 19 displays $E(R_n^+ \mid p)$ vs. p for $n = 5, 10, 15, 20$. For each n the mean is an increasing function in p and for each p is an increasing function in n ; in fact $E(R_n^+ \mid p) \doteq np$ for each n and p . Lemma 16 describes the relationship between the means at p and 1-p .

$\mathrm{Var}(R_n^+ \mid p)$ vs. p for $n = 5, 10, 15, 20$ is plotted in Figure 20. The variance is symmetric in p and 1-p , is zero for $p = 0$, and is maximized for $p = \frac{1}{2}$ for all n as is expected from the manner in which the probability masses shift as p varies. For fixed p the variance increases with n and for each n the variance increases as $p: 0 \longrightarrow \frac{1}{2}$.

It will be seen next that many of the properties of R_n^+ and the conditional random variable $\hat{\rho}_{n,\alpha}^+ = (R_n^+ \mid S_n^+ \geq s_{\alpha,p}^+)$ are similar. Note however that $R_n^+ \in \{0, \cdots, n\}$ whereas we would like $\hat{\rho}$ to satisfy $0 < \hat{\rho} < n$. We have assumed that a change in the model may have occurred only after sampling began, that is, $\rho > 0$, and ρ is estimated only when the null hypothesis that $\rho = n$ is rejected. Observe that the estimator $\hat{\rho}$ is consistent with both criteria: the largest value of S_n^+ for which $R_n^+ = 0, n$ is $s^+ = 1$; for each n and p, $P(S_n^+ \geq 1 \mid p)$ is very large and this s^+ will not be in the critical region for any reasonable size-α test. Hence $0 < \hat{\rho} < n$ as desired.

Each $j \in \{1, \cdots, n-1\}$ is a possible estimator for ρ when $\alpha \geq \sum_{k=1}^{n-1} p^k(1-p)^{n-k}$ since for $s^+ = 2$, $P((S_n^+, R_n^+) = (2, j) \mid p) = p^j(1-p)^{n-j}$. Thus all possible change points have positive probability of being detected when H_0 is rejected. As will be shown in Section 2.9 the next $\left[\frac{n}{2}\right]$ largest values attained by S_n^+ in decreasing order occur at j and $n-j$, $j=1, \cdots, \left[\frac{n}{2}\right]$, respectively.

In contrast to the estimator $\hat{\rho}$ for the test statistic S_n^+, consider the estimators for the change point for the generalized likelihook ratio test statistic (1.4.1) and Page's test statistic (1.4.2). For each of these statistics when the null hypothesis that no change has occurred is rejected, $\hat{\rho} = 0$ is a possible estimator for the change point; in fact for each test the most extreme values in the critical regions correspond to a change in the model occurring before sampling began. For the generalized likelihood ratio test each $\hat{\rho} \leq n-1$ is a possible estimator for ρ when p is small and q is large; when q is not large, however, recent change points cannot be detected. Moreover recent change points cannot be detected as p increases to larger values. Page's test statistic behaves similarly and for fixed p is even less able to detect changes late in the sequence than the generalized likelihood ratio test statistic.

Next the probability densities for $\hat{\rho}_{n,\alpha}^+$ are examined. The form of the densities are the same as those for R_n^+: for the same reasons it is also true for $\hat{\rho}_{n,\alpha}^+$ that when p is near 0 those points

near $j = 1$ are most probable and when p is near $\frac{1}{2}$ the extreme points are most probable; the densities for p and $1-p$ are related by $P(\hat{\rho}^+_{n,\alpha} = j \mid p) = P(\hat{\rho}^+_{n,\alpha} = n-j \mid 1-p)$. As a function of s^+, when s^+ decreases to the next smaller value, say $s^+_{\alpha,p'}$, then $P(\hat{\rho}^+_{n,\alpha} = j \mid p)$ increases for those points j at which $s^+_{\alpha,p}$ occurs and decreases for all other points j.

For $n = 10$ and $p = .1, .3, .5$, Table 21 lists $P(\hat{\rho}^+_{10,\alpha} = j \mid p)$ for $j = 1, \cdots, 9$, and the six largest values of S^+_{10} and the smallest value $\frac{22}{21}$ of S^+_{10} which is greater than 1; for each p the last line displays $P(R^+_{10} = j \mid p)$ for comparison. For many values of S^+_{10} and for p sufficiently small the more probable extreme points near $j = 1$ encompass even more of the probability mass for $\hat{\rho}^+_{10,\alpha}$ than for R^+_{10}.

The mean of $\hat{\rho}^+_{10,\alpha}$ is sketched in Figure 22 for $p \epsilon (0, \frac{1}{2}]$ and $s^+ = 2/1,\ 16/9,\ 8/5,\ 22/21$ and $0/1$. Lemma 16 describes the relationship between the means at p and $1-p$. For each s^+ the mean is tied down at $0, \frac{n}{2}$ when $p = 0, \frac{1}{2}$, respectively. As with R^+_{10}, $E(\hat{\rho}^+_{10,\alpha} \mid p)$ is an increasing function in p for each s^+. For $s^+ > 1$, $E(R^+_{10} \mid p,\ S^+_{10} \geq s^+) > E(R^+_{10} \mid p)$ when p is near 0 and $E(R^+_{10} \mid p,\ S^+_n \geq s^+) < E(R^+_{10} \mid p)$ when p is near $\frac{1}{2}$. Because of the nature of change in the density of $\hat{\rho}^+_{10,\alpha}$ as s^+ decreases, no monotonic properties as functions in s^+ can be stated for the mean of $\hat{\rho}^+_{10,\alpha}$ for arbitrary p.

Table 21. $P(\hat{\rho}^{+}_{10,\alpha} = j \mid p)$.

p=.1

s^+	j	α	j=1	j=2	j=3	j=4	j=5	j=6	j=7	j=8	j=9
2/1	1,...,9	.0436	.8889	.0988	.0110	.0012	.0001	2×10^{-5}	2×10^{-6}	2×10^{-7}	2×10^{-8}
16/9	1,9	.0780	.9379	.0552	.0061	.0007	7×10^{-5}	8×10^{-6}	9×10^{-7}	1×10^{-7}	8×10^{-7}
14/8	2,8	.0814	.8993	.0940	.0059	.0007	7×10^{-5}	8×10^{-6}	9×10^{-7}	6×10^{-6}	8×10^{-7}
12/7	3,7	.0817	.8958	.0938	.0098	.0007	7×10^{-5}	8×10^{-6}	5×10^{-5}	6×10^{-6}	8×10^{-7}
5/3	4,6	.0818	.8952	.0936	.0098	.0010	7×10^{-5}	.0003	5×10^{-5}	6×10^{-6}	8×10^{-7}
8/5	5	.0819	.8934	.0934	.0097	.0010	.0020	.0003	5×10^{-5}	6×10^{-6}	8×10^{-7}
22/21	3,7	.1099	.8163	.0810	.0692	.0073	.0146	.0027	.0055	.0018	.0027
0/1	0,2,4, 6,8,10	1.0000	.0897	.0721	.0078	.0112	.0152	.0159	.0049	.0079	.0112

p=.3

s^+	j	α	j=1	j=2	j=3	j=4	j=5	j=6	j=7	j=8	j=9
2/1	1,...,9	.0212	.5715	.2449	.1050	.0450	.0193	.0083	.0015	.0007	.0003
16/9	1,9	.0629	.8517	.0824	.0353	.0151	.0065	.0028	.0012	.0005	.0043
14/8	2,8	.0790	.6784	.2626	.0281	.0121	.0052	.0022	.0009	.0071	.0034
12/7	3,7	.0858	.6248	.2419	.0926	.0111	.0048	.0020	.0131	.0065	.0032
5/3	4,6	.0895	.5992	.2319	.0888	.0205	.0046	.0225	.0126	.0062	.0030
8/5	5	.0929	.5773	.2235	.0855	.0186	.0804	.0217	.0121	.0060	.0029
22/21	3,7	.3576	.5163	.1432	.0870	.0274	.0344	.0224	.0431	.0434	.0825
0/1	0,2,4, 6,8,10	1.0000	.1920	.1149	.0389	.0363	.0353	.0432	.0394	.0684	.1041

p=.5

s^+	j	α	j=1	j=2	j=3	j=4	j=5	j=6	j=7	j=8	j=9
2/1	1,...,9	.0088	.1111	.1111	.1111	.1111	.1111	.1111	.1111	.1111	.1111
16/9	1,9	.0244	.3600	.0400	.0400	.0400	.0400	.0400	.0400	.0400	.3600
14/8	2,8	.0381	.2308	.2051	.0256	.0256	.0256	.0256	.0256	.2051	.2308
12/7	3,7	.0498	.1765	.1569	.1373	.0196	.0196	.0196	.1373	.1569	.1765
5/3	4,6	.0586	.1500	.1333	.1167	.0917	.0167	.0917	.1167	.1333	.1500
8/5	5	.0645	.1364	.1212	.1061	.0833	.0909	.0833	.1061	.1212	.1364
22/21	3,7	.4873	.2686	.1123	.0702	.0320	.0341	.0320	.0702	.1123	.2686
0/1	0,2,4, 6,8,10	1.0000	.1826	.1100	.0547	.0449	.0342	.0449	.0547	.1100	.1826

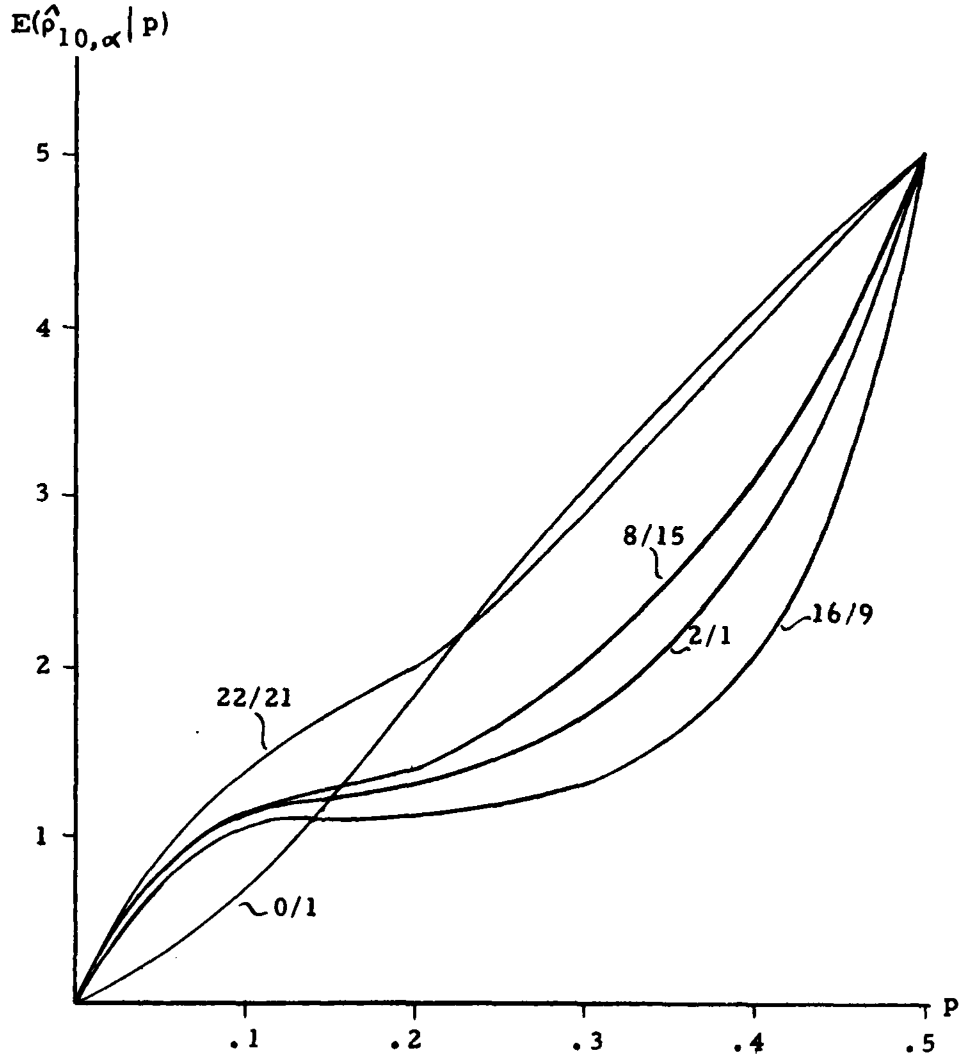


Figure 22. $E(\hat{\rho}^{+}_{10,\alpha}|p)$ vs. p .

Figure 23 displays the variance of $\hat{\rho}^+_{10,\alpha}$ for $p\epsilon(0,\frac{1}{2})$ and the same values of s^+. Symmetries occur in p and $1-p$, that is, $\mathrm{Var}(\hat{\rho}^+_{n,\alpha} \mid p) = \mathrm{Var}(\hat{\rho}^+_{n,\alpha} \mid 1-p)$. The variance is an increasing function in p for $p\epsilon\left[0,\frac{1}{2}\right]$ for each s^+. Again due to the nature of change in the densities as s^+ decreases no monotonic properties in s^+ can be stated for the variance of $\hat{\rho}^+_{10,\alpha}$ for arbitrary p. The tendency though is that as s^+ decreases to the next smaller value $s^+_{\alpha,p}$ then the variance decreases for p near 0 and increases for p near $\frac{1}{2}$ when $s^+_{\alpha,p}$ occurs nearer to $j = 1, n-1$ than the previous value of s^+. The opposite tendency occurs when $s^+_{\alpha,p}$ occurs nearer to $j = \left[\frac{n}{2}\right]$ than for the previous value of s^+.

This section is concluded by presenting an alternative to randomizing for estimating ρ when H_0 is rejected and S^+_n does not occur at a unique point. For sequences in the critical region $S^+_n(x_1,\cdots,x_n) > 1$ and it can be shown algebraically that if $J^+_n(x_1,\cdots,x_n) = \{j_1,\cdots,j_\ell\}$, $\ell \geq 2$, then $J^+_{n+1}(x_1,\cdots,x_n,x_{n+1})$ is single-valued. It is somewhat unexpected but in fact $J^+_{n+1}(x_1,\cdots,x_n,x_{n+1}) = \min\limits_{i=1,\cdots,\ell} j_i$ whether $x_{n+1} = 1$ or -1. Instead of randomizing then one might as well estimate ρ by $\min\limits_{i=1,\cdots,\ell} j_i$ and q using the last $n-\min\limits_{i=1,\cdots,\ell} j_i$ observations. For sequences with $S^+_n(x_1,\cdots,x_n) \leq 1$ and $J^+_n(x_1,\cdots,x_n) = \{j_1,\cdots,j_\ell\}$, $\ell \geq 2$, it is not true that $J^+_{n+1}(x_1,\cdots,x_n,x_{n+1}) = \min\limits_{i=1,\cdots,\ell} j_i$ independent of x_{n+1}.

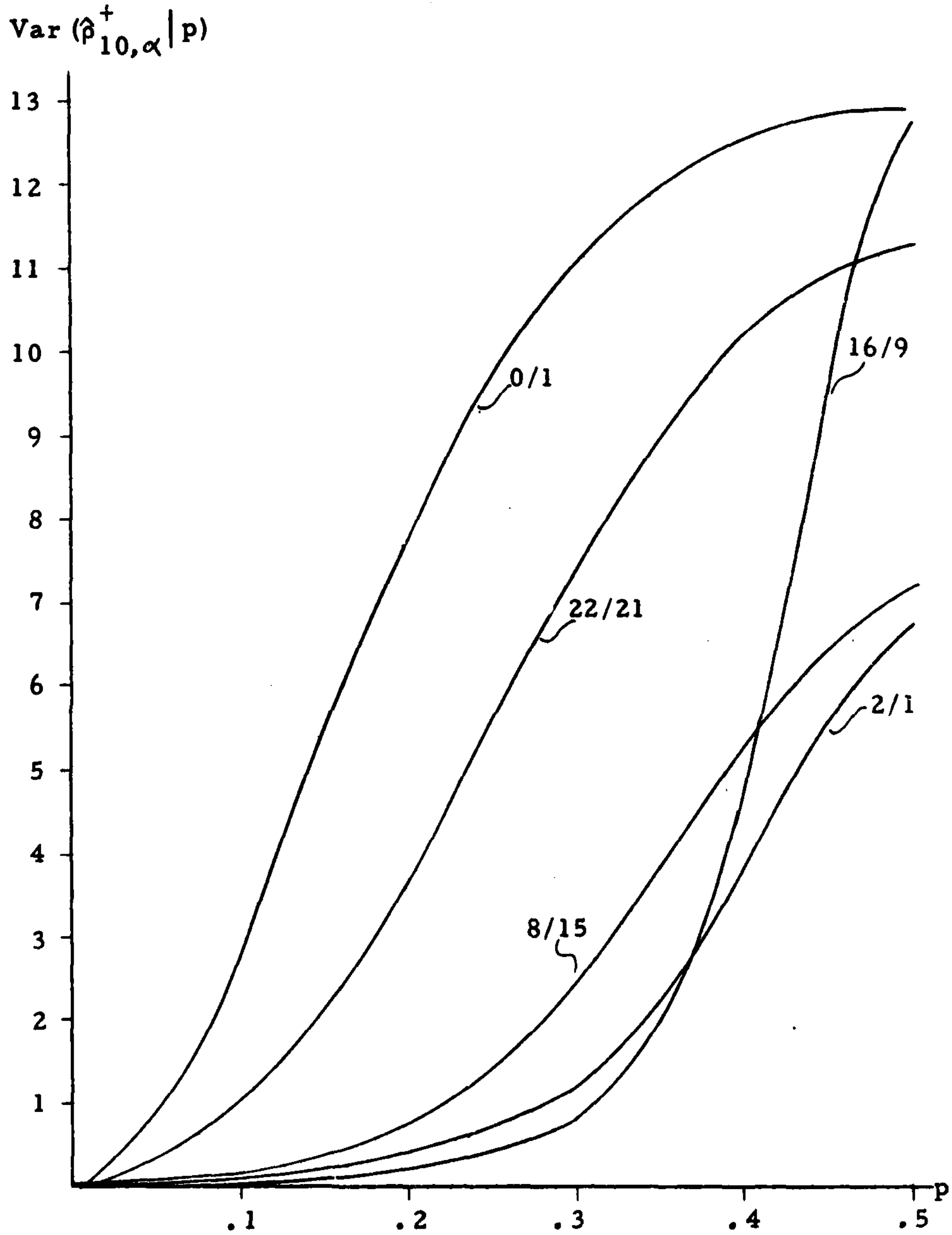


Figure 23. $Var(\hat{\rho}^{+}_{10,\alpha}|p)$ vs. p.

2.8 Properties of R_n

In this section the random variables R_n and $\hat{\rho}_{n,\alpha} = (R_n \mid S_n \geq s_{\alpha,p})$ are studied. The density of R_n and its mean and variance are first examined; comparisons are made between R_n and R_n^+. Next $\hat{\rho}_{n,\alpha}$ is investigated; it will be seen that R_n and $\hat{\rho}_{n,\alpha}$ behave similarly.

The densities of R_n are sketched in Figures 24.1, 24.2 for $n = 10, 20$ and $p = .1, .2, .3, .4, .5$; in contrast to the densities for R_n^+ note that the densities for R_n are symmetrical in j and $n-j$ and $P(R_n = j \mid p) = P(R_n = j \mid 1-p)$ for all n, p and j by Theorem 6. Observe the following behavior of the densities as suggested by theory and computation.

The sequences for which $R_n = 0, n$ are of the types $T_n^k(0,0)$, $T_n^k(n,k)$ where $k, n-k \leq \left[\frac{n-1}{4}\right]$ if n is even or $k, n-k < \left[\frac{n-2}{4}\right]$ if n is odd by Theorem 7. For these types of sequences the tendency is for $P(R_n = 0 \mid p)$ to decrease as $p: 0 \longrightarrow \frac{1}{2}$. For each p, $P(R_n = 0 \mid p)$ tends to decrease as n increases and the proportion of sequences with $R_n = 0$ is decreasing with n. Observe that $P(R_n = 0 \mid p) < P(R_n^+ = 0 \mid p)$ for $p \epsilon (0, \frac{1}{2})$; this is expected since $S_n = \max \{S_n^+, -S_n^+\}$.

For $j \epsilon \left\{1, \cdots, \left[\frac{n-2}{2}\right]\right\}$ it appears to be the case that $\sum_{k=0}^{n} d_k(j) \geq \sum_{k=0}^{n} d_k(j+1)$; in general those sequences for which $R_n = j$ either have a larger proportion of 1's than -1's in $x_1, \cdots, x_j$ and a larger

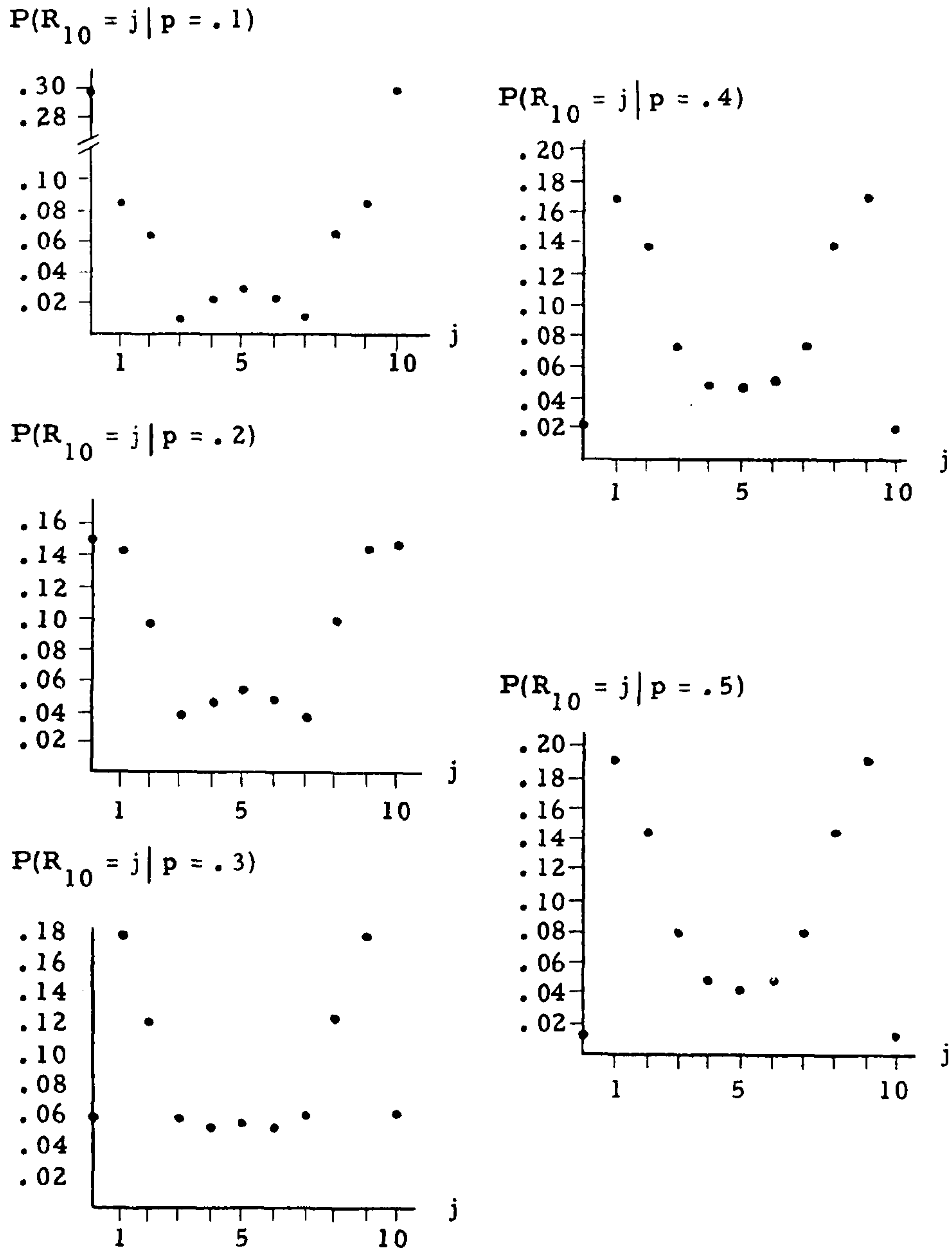


Figure 24.1 $P(R_{10} = j \mid p)$ vs. j .

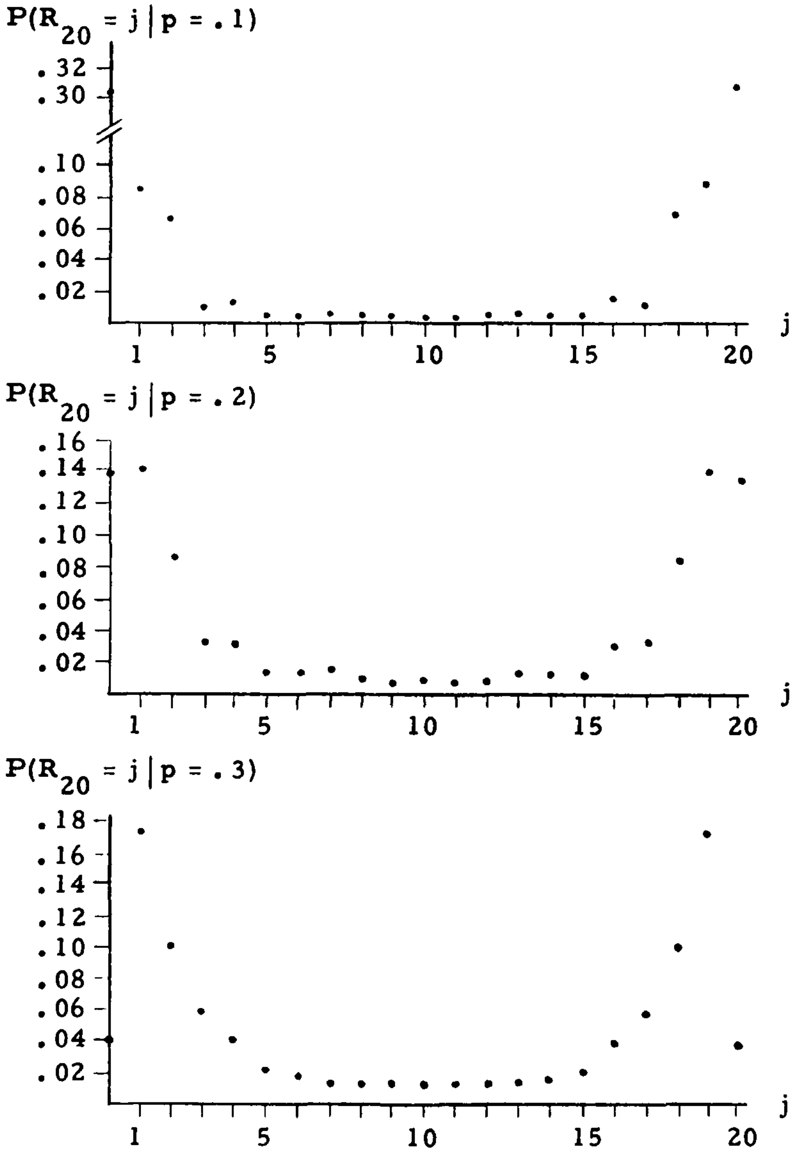
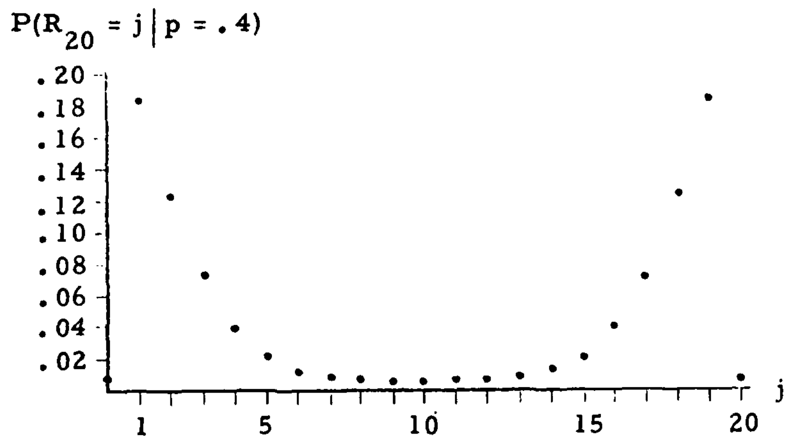


Figure 24.2 $P(R_{20} = j \mid p)$ vs. j .



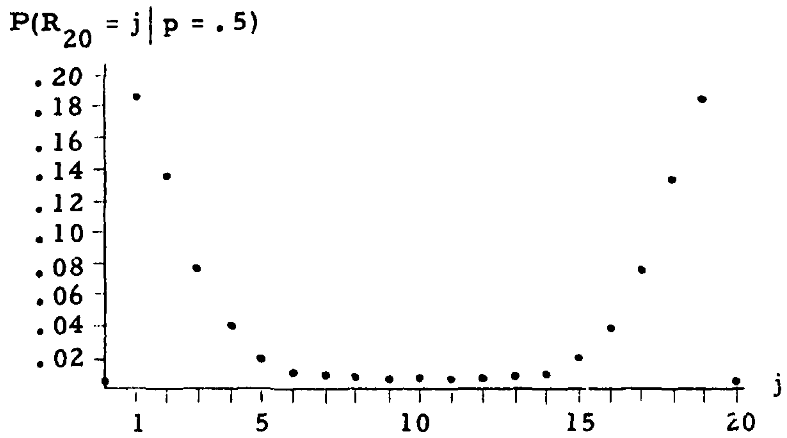


Figure 24.2 (Continued)

proportion of -1's than 1's in $x_{j+1}, \cdots, x_n$ or vice versa. Thus the extreme points near $j = 1, n-1$ are most probable. Also calculations indicate that for the more probable extreme points j, $P(R_{n_1} = j \mid p) \doteq P(R_{n_2} = j \mid p)$ for each n_1, n_2 and p. The same phenomenon occurs for R_n^+. Furthermore note that for each n, $p \epsilon (0, \frac{1}{2})$ and $j \leq \left[\frac{n}{2}\right]$, $P(R_n^+ = j \mid p) \doteq P(R_n = j \mid p)$, as is anticipated from the manner in which S_n^+ and S_n are defined.

Due to the symmetry in j and $n-j$ for the density of R_n, $E(R_n \mid p) = \frac{n}{2}$ and $\mathrm{Var}(R_n \mid p) = \mathrm{Var}(R_n \mid 1-p)$ for all n and p. Figure 25 displays $\mathrm{Var}(R_n \mid p)$ for $n = 5, 10, 15, 20$. Since opposite extreme points encompass most of the probability mass the variance is quite large. The variance is maximized for $p = 0$ where $\mathrm{Var}(R_n \mid p = 0) = \frac{n^2}{4}$; in general the variance decreases as $p\colon 0 \longrightarrow \frac{1}{2}$ for each n and increases as n increases for each p. In comparison with R_n^+, $\mathrm{Var}(R_n^+ \mid p) \doteq \mathrm{Var}(R_n \mid p)$ for $.35 < p < .65$ and $\mathrm{Var}(R_n^+ \mid p) << \mathrm{Var}(R_n \mid p)$ for p near $0, 1$.

The conditional random variable $\hat{\rho}_{n,\alpha}$ is studied next. Note that $0 < \hat{\rho}_{n,\alpha} < n$ for all reasonable size-α tests and each point $j \epsilon \{1, \cdots, n-1\}$ is a possible estimator for ρ when the null hypothesis is rejected.

The densities of $\hat{\rho}_{10,\alpha}$ for $p = .1, .3, .5$ and for the six largest values of S_{10} and the smallest value of S_{10} which is greater than 1 are listed in Table 26; for each p the last line lists

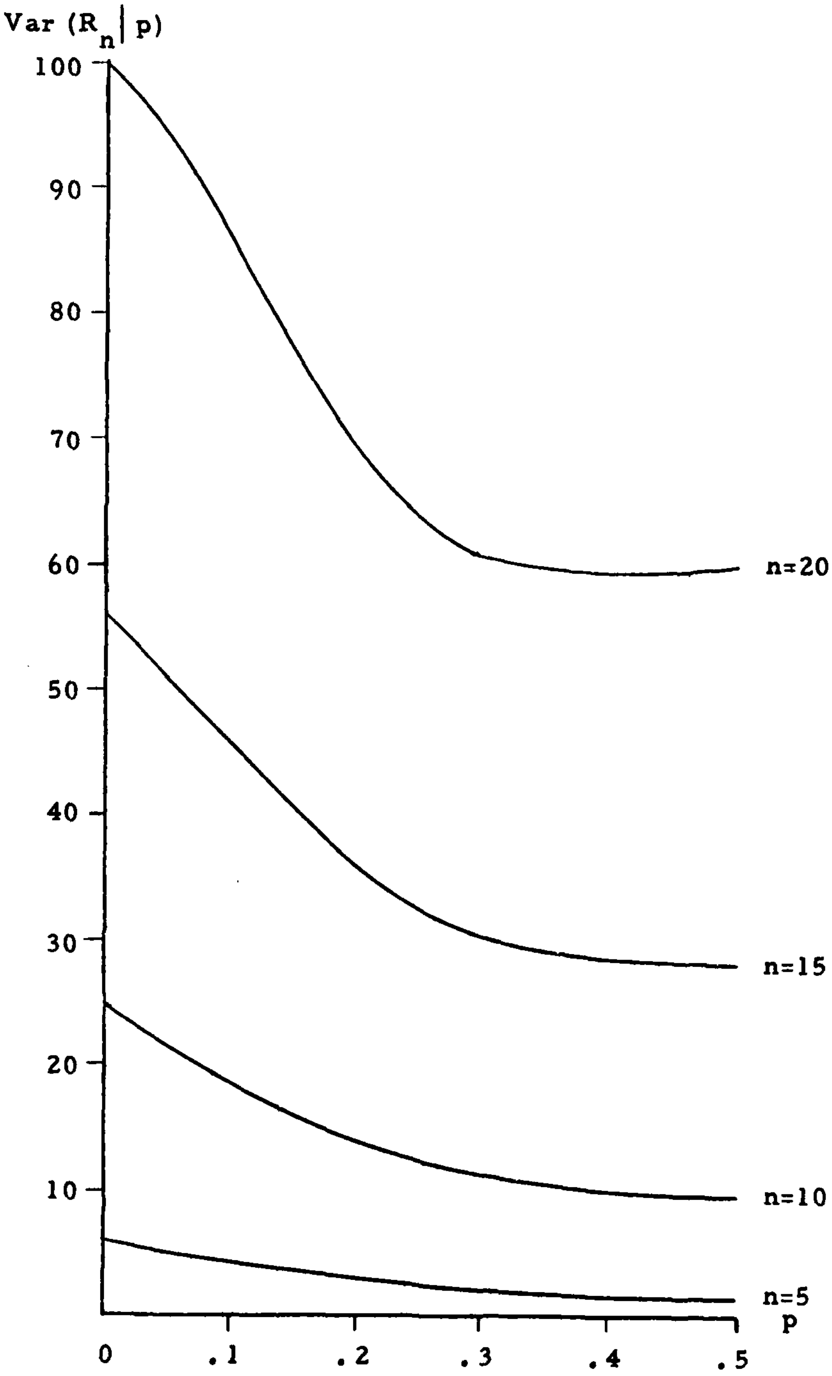


Figure 25. Var $(R_n|p)$ vs. p .

Table 26. $P(\hat{\rho}_{10,\alpha} = j \mid p)$.

p=.1

s	j	α	j=1	j=2	j=3	j=4	j=5
2/1	1,...,9	.0872	.4444	.0494	.0055	.0006	.0001
16/9	1,9	.1517	.4681	.0284	.0032	.0004	.0001
14/8	2,8	.1588	.4483	.0483	.0030	.0003	.0001
12/7	3,7	.1591	.4465	.0481	.0050	.0003	.0001
5/3	4,6	.1592	.4462	.0481	.0050	.0007	.0001
8/5	5	.1595	.4453	.0480	.0050	.0007	.0021
22/21	3,7	.2083	.4075	.0432	.0365	.0048	.0158
6/10	0,10	1.0000	.0860	.0663	.0110	.0230	.0291

p=.3

s	j	α	j=1	j=2	j=3	j=4	j=5
2/1	1,...,9	.0424	.2862	.1233	.0543	.0266	.0193
16/9	1,9	.1207	.4249	.0433	.0191	.0093	.0068
14/8	2,8	.1528	.3355	.1394	.0150	.0074	.0053
12/7	3,7	.1664	.3082	.1281	.0545	.0068	.0049
5/3	4,6	.1737	.2951	.1227	.0522	.0277	.0047
8/5	5	.1805	.2841	.1181	.0502	.0266	.0420
22/21	3,7	.6054	.2747	.1052	.0722	.0276	.0406
6/10	0,10	1.0000	.1774	.1214	.0598	.0524	.0560

p=.5

s	j	α	j=1	j=2	j=3	j=4	j=5
2/1	1,...,9	.0176	.1111	.1111	.1111	.1111	.1111
16/9	1,9	.0469	.3542	.0417	.0417	.0417	.0417
14/8	2,8	.0742	.2237	.2105	.0263	.0263	.0263
12/7	3,7	.0977	.1700	.1600	.1400	.0202	.0202
5/3	4,6	.1152	.1441	.1356	.1186	.0932	.0169
8/5	5	.1270	.1308	.1231	.1077	.0846	.1077
22/21	3,7	.7813	.2263	.1312	.0837	.0375	.0425
6/10	0,10	1.0000	.1927	.1465	.0791	.0485	.0449

$P(R_{10} = j \mid p)$ for comparison. The densities for $\hat{\rho}_{10,\alpha}$ are similar to those for R_{10} in that the extreme points are most probable; for p near 0 the extreme points $j = 1, 9$ are more probable for $\hat{\rho}_{10,\alpha}$ than for R_{10}.

The mean of $\hat{\rho}_{n,\alpha}$ is $\frac{n}{2}$ for each $n, p\epsilon(0, 1), \alpha$. The variance of the conditional random variable is sketched in Figure 27 for $p\epsilon(.1, .5)$ and $s^+ = 2/1, 16/9, 8/5, 22/21, 6/10$; symmetries occur in p and 1-p . Similar to the behavior of the variance of R_{10}, the conditional variance tends to be large for p small and to decrease as p increases towards $\frac{1}{2}$; for p sufficiently small $\text{Var}(\hat{\rho}_{10,\alpha} \mid p$ $\text{Var}(R_{10} \mid p)$. Note however that $\text{Var}(R_{10} \mid p)$ is maximized for $p = 0$ whereas $\text{Var}(\hat{\rho}_{10,\alpha} \mid p = 0) = 0$ for $S_n > 1$. Due to the nature of the densities of $\hat{\rho}_{10,\alpha}$ the tendency is that as s decreases to the next smaller value $s_{\alpha,p}$; the conditional variance increases when $s_{\alpha,p}$ occurs nearer to $j = 1, n-1$ than the previous value of s . For p in a neighborhood of $\frac{1}{2}$, $\text{Var}(\hat{\rho}^+_{10,\alpha} \mid p) \doteq \text{Var}(\hat{\rho}_{10,\alpha} \mid p)$ and $\text{Var}(\hat{\rho}^+_{10,\alpha} \mid p) << \text{Var}(\hat{\rho}_{10,\alpha} \mid p)$ for p nearer to 0, 1 .

2.9 Large Sample Derivations

Expressions defining critical regions containing more than $\left[\frac{n}{2}\right] + 1$ values of the test statistics appear to be intangible for arbitrary n . The reason is apparent from the procedure for counting employed in Section 2.2. Although the values attained by the statistics as functions in n are available from the lists constructed in I. of the

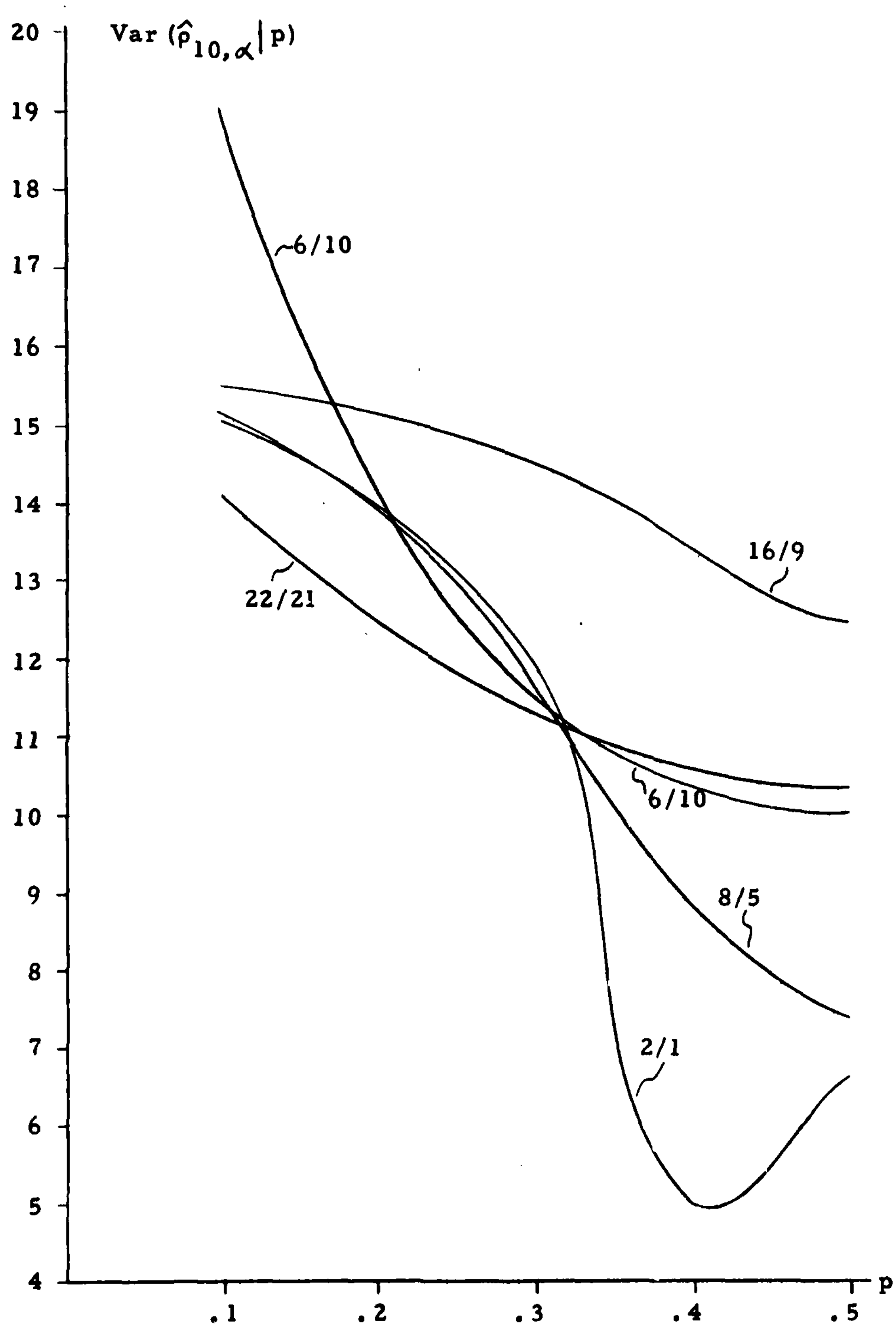


Figure 27. Var$(\hat{\rho}_{10,\alpha}|p)$ vs. p .

Table 28. Large sample derivations.

s^+	j	$T_n^k(r, a)$	$c_k^+(s^+, j)$	
$\frac{2}{1}$	1	$T_n^1(1, 1)$	1	
	2	$T_n^2(2, 2)$	1	
	$\vdots$	$\vdots$	$\vdots$	
	$n-1$	$T_n^{n-1}(n-1, n-1)$	1	
$\frac{2(n-i-1)}{n-i}$	i	$T_n^{i+1}(i, i)$	$n-i-1$	for $i=1, \ldots, \left[\frac{n}{2}\right] - 2$
	$n-i$	$T_n^{n-i-1}(n-i, n-i-1)$	$n-i-1$	
$\frac{2n}{n+2}$	$\frac{n-2}{2}$	$T_n^{\frac{n}{2}}(\frac{n-2}{2}, \frac{n-2}{2})$	$\frac{n-1}{2}$	if n is even
	$\frac{n+2}{2}$	$T_n^{\frac{n}{2}}(\frac{n+2}{2}, \frac{n}{2})$	$\frac{n-1}{2}$	

Table 28. (Continued).

s^+	j	$T_n^k(r, a)$	$c_k^+(s^+, j)$	
$\frac{2(n-2)}{n}$	$\frac{n}{2}$	$T_n^{\frac{n+2}{2}}(\frac{n}{2}, \frac{n}{2})$	$\frac{n-4}{2}$	if n is even
	$\frac{n}{2}$	$T_n^{\frac{n-2}{2}}(\frac{n}{2}, \frac{n-2}{2})$	$\frac{n-4}{2}$	
$\frac{2(n-\left[\frac{n}{2}\right])}{n-\left[\frac{n}{2}\right]+1}$	$\left[\frac{n}{2}\right] - 1$	$T_n^{\left[\frac{n}{2}\right]}(\left[\frac{n}{2}\right]-1, \left[\frac{n}{2}\right]-1)$	$n - \left[\frac{n}{2}\right]$	if n is odd
	$n-\left[\frac{n}{2}\right]+1$	$T_n^{\left[\frac{n}{2}\right]+1}(n-\left[\frac{n}{2}\right]+1, \left[\frac{n}{2}\right]+1$	$n - \left[\frac{n}{2}\right]$	
$\frac{2(n-\left[\frac{n}{2}\right]-1)}{n-\left[\frac{n}{2}\right]}$	$\left[\frac{n}{2}\right]+1$	$T_n^{n-\left[\frac{n}{2}\right]}(\left[\frac{n}{2}\right], \left[\frac{n}{2}\right])$	$\left[\frac{n}{2}\right]-1$	
	$\left[\frac{n}{2}\right]$	$T_n^{\left[\frac{n}{2}\right]+1}(\left[\frac{n}{2}\right]+1, 1)$	$\left[\frac{n}{2}\right]-1$	

Table 28. (Continued).

s	j	$T_n^k(r, a)$	$c_k(s, j)$	
$\frac{2}{1}$	1	$T_n^1(1, 1),\ T_n^{n-1}(1, 0)$	$1, 1$	
	2	$T_n^2(2, 2),\ T_n^{n-2}(2, 0)$	$1, 1$	
	$\vdots$	$\vdots$	$\vdots$	
	$n-1$	$T_n^{n-1}(n-1, n-1),\ T_n^1(n-1, 0)$	$1, 1$	
$\frac{2(n-2)}{n-1}$	1	$T_n^2(1, 1),\ T_n^{n-2}(1, 0)$	$\frac{2n-5}{2},\ \frac{2n-5}{2}$	
	$n-1$	$T_n^{n-2}(n-1, n-2),\ T_n^2(n-1, 1)$	$\frac{2n-5}{2},\ \frac{2n-5}{2}$	
$\frac{2(n-i-1)}{n-i}$	i	$T_n^{i+1}(i, i),\ T_n^{n-i-1}(i, 0)$	$n-i-1,\ n-i-1$	for $i=2,\cdots,\left[\frac{n}{2}\right]-2$
	$n-i$	$T_n^{n-i}(n-i, n-i-1),\ T_n^{i+1}(n-i, 1)$	$n-i-1,\ n-i-1$	
$\frac{2n}{n+2}$	$\frac{n-2}{2}$	$T_n^{\frac{n}{2}}(\frac{n-2}{2}\ \ \frac{n-2}{2}),\ T_n^{\frac{n}{2}}(\frac{n-2}{2}, 0)$	$\frac{n-1}{2},\ \frac{n-1}{2}$	if n is even
	$\frac{n+2}{2}$	$T_n^{\frac{n}{2}}(\frac{n+2}{2}, \frac{n}{2}),\ T_n^{\frac{n}{2}}(\frac{n+2}{2}, 1)$	$\frac{n-1}{2},\ \frac{n-1}{2}$	

Table 28. (Continued).

s	j	$T_n^k(r, a)$	$c_k(s, j)$	
$\frac{2(n-2)}{n}$	$\frac{n}{2}$	$T_n^{\frac{n+2}{2}}(\frac{n}{2}, \frac{n}{2})$, $T_n^{\frac{n-2}{2}}(\frac{n}{2}, 0)$	$\frac{n-4}{2}$, $\frac{n-4}{2}$	if n is even
	$\frac{n}{2}$	$T_n^{\frac{n-2}{2}}(\frac{n}{2}, \frac{n-2}{2})$, $T_n^{\frac{n+2}{2}}(\frac{n}{2}, 1)$	$\frac{n-4}{2}$, $\frac{n-4}{2}$	
$\frac{2(n-\left[\frac{n}{2}\right])}{n-\left[\frac{n}{2}\right]+1}$	$\left[\frac{n}{2}\right]-1$	$T_n^{\left[\frac{n}{2}\right]}(\left[\frac{n}{2}\right]-1, \left[\frac{n}{2}\right]-1)$, $T_n^{n-\left[\frac{n}{2}\right]}(\left[\frac{n}{2}\right]-1, 0)$	$n-\left[\frac{n}{2}\right]$, $n-\left[\frac{n}{2}\right]$	if n is odd
	$n-\left[\frac{n}{2}\right]+1$	$T_n^{n-\left[\frac{n}{2}\right]+1}(n-\left[\frac{n}{2}\right]+1, n-\left[\frac{n}{2}\right])$, $T_n^{\left[\frac{n}{2}\right]}(n-\left[\frac{n}{2}\right]+1, 1)$	$n-\left[\frac{n}{2}\right]$, $n-\left[\frac{n}{2}\right]$	
$\frac{2(n-\left[\frac{n}{2}\right]-1)}{n-\left[\frac{n}{2}\right]}$	$\left[\frac{n}{2}\right]$	$T_n^{\left[\frac{n}{2}\right]+1}(\left[\frac{n}{2}\right], \left[\frac{n}{2}\right])$, $T_n^{\left[\frac{n}{2}\right]}(\left[\frac{n}{2}\right], 0)$	$\left[\frac{n}{2}\right]-1$, $\left[\frac{n}{2}\right]-1$	
	$\left[\frac{n}{2}\right]+1$	$T_n^{\left[\frac{n}{2}\right]}(\left[\frac{n}{2}\right]+1, \left[\frac{n}{2}\right])$, $T_n^{\left[\frac{n}{2}\right]+1}(\left[\frac{n}{2}\right]+1, 1)$	$\left[\frac{n}{2}\right]-1$, $\left[\frac{n}{2}\right]-1$	

procedure, one must then be able to order at least the larger values of the statistics within each category of a fixed number of -1's in the sequence to count the $c_k^+(s^+, j)$'s and $c_k(s, j)$'s as in III. This ordering cannot be accomplished for arbitrary n except for the most extreme values. Table 28 lists the $\left[\frac{n}{2}\right]+1$ largest values of the test statistics and other pertinent information needed for calculating probabilities. For each p and large n the critical regions containing only these values will have very small probability.

Even though larger exact critical regions for the tests for arbitrary n cannot be derived, observe from Table 9 that for each $n \leq 20$, $P(S_n^+ > 1 \mid p = .1) \doteq .108$ and similar but less emphatic statements can be made for $p > .1$. It appears as though approximations to critical regions may be obtainable for any n and p, but the method of proof for establishing stability properties and providing bounds in the error in approximation is evasive at present.

2.10 Example

The following example illustrates a change point testing situation.

Measurements of the discharge of the Missouri River at the Omaha Station are taken throughout the year with the peak discharge occurring in spring. This greatest flow may vary considerably from year to year: when snowfall has been heavy and spring thaw quick the peak discharge may be quite large. It is the peak discharge that is of

interest for flood control purposes. We test for a downward change in the peak discharge during the interval of years 1945-1964.

Table 29 lists the peak discharges $y_1, \cdots, y_{20}$ for the years 1945-1964. Since the data are so variable the median of the peak discharge is used as a measure of central location for constructing the test. An estimate $\hat{m}$ of the median under the null hypothesis is found from the peak discharges for the years 1930-1940 as listed in the same table; we find $\hat{m} = 85,600$ cubic feet/second. Define

$$x_i = \begin{cases} 1 & \text{if } y_i \geq 85,600 \\ -1 & \text{if } y_i < 85,600 \end{cases}, \quad i = 1, \cdots, 20,$$ and for these transformed data

test H_0: $\rho = 20,\ p = \frac{1}{2}$ vs. H_1^+: $\rho < 20,\ p = \frac{1}{2},\ q,\ p > q$. Using the x_i's as computed in Table 29 we find that $S_{20}^+ = \frac{140}{99}$ and $R_{20}^+ = 9$. Now $P(S_{20}^+ \geq \frac{140}{99} \mid p = \frac{1}{2}) = .057$ and when H_0 is rejected at this significance level, a decrease in the median peak discharge is estimated to have occurred in 1954. River flow controls were incidently instigated upstream from Omaha at this time; the change point test detected a simultaneous decrease in peak discharge.

Note that the change point test detects a decrease in peak discharge of the river with fewer than 20 observations. If for each set of 15 consecutive years between 1945 and 1964 one calculates S_{15}^+, one finds that $S_{15}^+(x_1, \cdots, x_{15}) = \frac{16}{9}$, $S_{15}^+(x_2, \cdots, x_{16}) = \frac{41}{28}$, $S_{15}^+(x_3, \cdots, x_{17}) = \frac{7}{4}$, $S_{15}^+(x_4, \cdots, x_{18}) = \frac{14}{9}$, $S_{15}^+(x_5, \cdots, x_{19}) = \frac{8}{5}$ and $S_{15}^+(x_6, \cdots, x_{20}) = \frac{18}{11}$; the probability of each value being at

Table 29. Peak Discharge y_i in Cubic Feet/Second of the Missouri River at the Omaha Station.

Year	y_i	Year	y_i	x_i
1930	62,400	1945	106,000	1
1931	45,000	1946	81,900	-1
1932	104,000	1947	147,000	1
1933	89,700	1948	97,100	1
1934	32,900	1949	169,000	1
1935	85,600	1950	194,000	1
1936	78,500	1951	86,200	1
1937	110,000	1952	536,000	1
1938	114,000	1953	111,000	1
1939	140,000	1954	80,300	-1
1940	34,300	1955	34,700	-1
		1956	36,900	-1
		1957	38,000	-1
		1958	31,600	-1
		1959	37,100	-1
		1960	117,000	1
		1961	30,300	-1
		1962	103,000	1
		1963	32,700	-1
		1964	35,700	-1

least so large is .004, .067, .004, .041, .022 and .019, respectively. For each set of observations R^+_{15} is the point corresponding to the year 1953; rejecting the null hypothesis results in estimating a decrease in greatest annual flow beginning in 1954.

Similarly for each set of 10 consecutive years one calculates for S^+_{10} the values $\frac{16}{9}$, $\frac{7}{4}$, 2, 2, 2, 2, $\frac{12}{7}$, $\frac{7}{4}$, $\frac{14}{9}$, $\frac{6}{10}$ and $\frac{6}{10}$; the probability of each value being at least so large is .024, .038, .008, .008, .008, .008, .049, .038, .119, .797 and .797 , respectively. For the first nine values R^+_{10} is the point corresponding to the year 1953; for the last two values $R^+_{10}(x_{10}, \cdots, x_{19}) = R^+_{10}(x_{11}, \cdots, x_{20}) = 0$.

Also for each set of 5 consecutive years between 1950 and 1957, $S^+_5 = 2$ and $P(S^+_5 = 2 \mid p = \frac{1}{2}) = .125$. For these sets of observations R^+_5 corresponds to the year 1953; rejecting the null hypothesis for size of test $\alpha = .125$ results in estimating a decrease in peak discharge beginning in 1954. For each other set of 5 consecutive years $S^+_5 < 2$.

Next we investigate whether Page's test statistic also detects a change in peak discharge. For the one-sided test with $p = \frac{1}{2} > q$ Page's test statistic is $M = \min_{0 \le r \le n} \left[S_r - \max_{0 \le i < r} S_i \right]$ where $S_r = \sum_{i=1}^{r} X_i$, $r = 1, \cdots, n$, and $S_0 = 0$. For $n = 20, 15, 10$ and $\alpha = .05$, the critical regions contain those sequences for which $M \le -10, -9, -7$, respectively. Critical regions for these values of n and α were found by

using tables constructed by Page (1955), performing calculations for smaller n than listed in the tables and interpolating. For the transformed data $x_1, \cdots, x_{20}$, $M = -7$; for each set of 15 consecutive years between 1945 and 1964, $M \geq -7$; and for each set of 10 consecutive years, $M \geq -6$. Hence in none of these instances does Page's test detect a change for size of test $\alpha = .05$. For $n = 5$ and $\alpha = .05$, the critical region contains the sequence -1-1-1-1-1. For the two sets of years 1954-1958 and 1955-1959, Page's test detects a change in peak discharge for this size of test and the change is estimated to have occurred before sampling began.

CHAPTER 3

ALTERNATIVE HYPOTHESIS DISTRIBUTIONS

The alternative hypothesis distributions of the test statistics and the change point random variables are found from the joint distributions $(S_n^+, R_n^+ \mid p, \rho, q)$ and $(S_n, R_n \mid p, \rho, q)$. Obtaining expressions for the joint distributions is an incredibly tedious task; results facilitating this task are presented in Section 1 and application of these results is illustrated in Section 2. Properties of the point distributions are derived in the next two sections; from these properties the behavior of the marginal distributions of the test statistics and estimators of change points are deduced.

3.1 Notation and Tools

Let $c_k(s, j \mid \rho, a)$ denote the number of sequences $x_1, \cdots, x_n$ with k 1's and $n-k$ -1's that have $S_n(x_1, \cdots, x_n) = s$, $R_n(x_1, \cdots, x_n) = j$ and a 1's in the first ρ places. Thus

3.1.1 $$P\{(S_n, R_n) = (s, j) \mid p, \rho, q\}$$

$$= \sum_{k=0}^{n} \sum_{a=0}^{\rho} c_k(s, j \mid \rho, a)\, p^a (1-p)^{\rho-a} q^{k-a} (1-q)^{n-\rho-(k-a)} .$$

Define $c_k(s \mid \rho, a) = \sum_{j=0}^{n} c_k(s, j \mid \rho, a)$ so that

3.1.2 $$P(S_n = s \mid p, \rho, q) = \sum_{k=0}^{n} \sum_{a=0}^{\rho} c_k(s \mid \rho, a)\, p^a (1-p)^{\rho - a} \cdot q^{k-a} (1-q)^{n-\rho-(k-a)}$$

and $C_k(s_{\alpha,p} \mid \rho, a) = \sum_{s \geq s_{\alpha,p}} c_k(s \mid \rho, a)$ so that

3.1.3 $$P(S_n \geq s_{\alpha,p} \mid p, \rho, q) = \sum_{k=0}^{n} \sum_{a=0}^{\rho} C_k(s_{\alpha,p} \mid \rho, a)\, p^a (1-p)^{\rho-a} \cdot q^{k-a} (1-q)^{n-\rho-(k-a)} .$$

Define $d_k(j \mid \rho, a, s_{\alpha,p}) = \sum_{s \geq s_{\alpha,p}} c_k(s, j \mid \rho, a)$ so that

3.1.4 $$P(\hat{\rho}_{n,\alpha} = j \mid p, \rho, q) = \frac{\sum_{k=0}^{n} \sum_{a=0}^{\rho} d_k(j \mid \rho, a, s_{\alpha,p})\, p^a (1-p)^{\rho-a} q^{k-a} (1-q)^{n-\rho-(k-a)}}{P(S_n \geq s_{\alpha,p} \mid p, \rho, q)} .$$

Similarly define $c_k^+(s^+, j \mid \rho, a)$, $c_k^+(s^+ \mid \rho, a)$, $C_k^+(s^+_{\alpha,p} \mid \rho, a)$ and $d_k^+(j \mid \rho, a, s^+_{\alpha,p})$.

The $c_k^+(s^+, j \mid \rho, a)$'s and $c_k(s, j \mid \rho, a)$'s must be counted to find the joint distributions under the alternative hypothesis. Results useful for minimizing the effort needed to count these quantities are now presented. The one-sided testing situation is considered first.

<u>Theorem 1</u>. For each n, k, s^+, j, ρ and a, $c_k^+(s^+, j \mid \rho, a)$ $= c_{n-k}^+(s^+, n-j \mid n-\rho, n-\rho-a(k-1))$; hence $P\{(S_n^+, R_n^+)=(s^+, j) \mid p, \rho, q\}$ $= P\{(S_n^+, R_n^+) = (s^+, n-j) \mid 1-q, n-\rho, 1-p\}$ for all p and q.

<u>Proof.</u> If $S_n^+(x_1, \cdots, x_n)$ occurs at j and if $x_1, \cdots, x_n$ has k 1's with a 1's in the first ρ places, then $S_n^+(-x_n, \cdots, -x_1)$ occurs at $n-j$ and $-x_n, \cdots, -x_1$ has $n-k$ 1's with $n-\rho-(k-a)$ 1's in the first $n-\rho$ places. ■

Hence the $c_k^+(s^+, j \mid \rho, a)$'s need only be counted for $k \leq \left[\frac{n}{2}\right]$ or for $\rho \leq \left[\frac{n}{2}\right]$ or for $j \leq \left[\frac{n}{2}\right]$. Numerical computation need only be carried out for $j \leq \left[\frac{n}{2}\right]$ or for $p \in (0, \frac{1}{2}]$ to evaluate the joint probabilities.

<u>Corollary 2.</u> (i) For each n, k, s^+, ρ and a, $c_k^+(s^+ \mid \rho, a)$ $= c_{n-k}^+(s^+ \mid n-\rho, n-\rho-(k-a))$ and $C_k^+(s^+ \mid \rho, a) = C_{n-k}^+(s^+ \mid n-\rho, n-\rho-(k-a))$. Thus $P(S_n^+ \geq s^+ \mid p, \rho, q) = P(S_n^+ \geq s^+ \mid 1-q, n-\rho, 1-p)$ for all p and q.

(ii) For each n, k, j, ρ, a and $s_{\alpha,p}^+$, $d_k^+(j \mid \rho, a, s_{\alpha,p}^+)$ $= d_{n-k}^+(n-j \mid n-\rho, n-\rho-(k-a), s_{\alpha,p}^+)$. Thus $P(\hat{\rho}_{n,\alpha}^+ = j \mid p, \rho, q)$ $= P(\hat{\rho}_{n,\alpha}^+ = n-j \mid 1-p, n-\rho, 1-q)$ for all p and q.

As a check when counting note that $\sum_{a=0}^{\rho} c_k^+(s^+, j \mid \rho, a) =$ $c_k^+(s^+, j)$, $\sum_{a=0}^{\rho} c_k^+(s^+ \mid \rho, a) = c_k^+(s^+)$ and $\sum_{a=0}^{\rho} d_k^+(j \mid \rho, a, s^+_{\alpha, p})$ $= d_k^+(j \mid s^+_{\alpha, p})$ for each k, s^+, j, ρ, a and $s^+_{\alpha, p}$.

Similar results hold for the two-sided testing situation.

Theorem 3. For each n, k, s, j, ρ and a, $c_k(s, j \mid \rho, a) = c_k(s, n-j \mid n-\rho, k-a) = c_{n-k}(s, j \mid \rho, \rho-a) = c_{n-k}(s, n-j \mid n-\rho, n-\rho-(k-a))$; hence $P\{(S_n, R_n) = (s, j) \mid p, \rho, q\} = P\{(S_n, R_n) = (s, n-j) \mid q, n-\rho, p\} = P\{(S_n, R_n) = (s, j) \mid 1-p, \rho, 1-q\} = P\{(S_n, R_n) = (s, n-j) \mid 1-q, n-\rho, 1-p\}$ for all p and q.

Proof. If $S_n(x_1, \cdots, x_n)$ occurs at j and if $x_1, \cdots, x_n$ has k 1's with a 1's in the first ρ places, then $S_n(x_n, \cdots, x_1)$ occurs at $n-j$ and $x_n, \cdots, x_1$ has k 1's with $k-a$ 1's in the first $n-\rho$ places, and $S_n(-x_1, \cdots, -x_n)$ occurs at j and $-x_1, \cdots, -x_n$ has $n-k$ -1's with $\rho-a$ 1's in the first ρ places. ∎

Hence the $c_k(s, j \mid \rho, a)$'s need only be counted for $k \leq \left[\frac{n}{2}\right]$ and $j \leq \left[\frac{n}{2}\right]$ or for $k \leq \left[\frac{n}{2}\right]$ and $\rho \leq \left[\frac{n}{2}\right]$. Numerical computation need only be carried out for $j \leq \left[\frac{n}{2}\right]$ and $p \in (0, \frac{1}{2}]$.

Corollary 4. (i) For each n, k, s, ρ and a, $c_k(s \mid \rho, a) = c_k(s \mid n-\rho, k-a) = c_{n-k}(s \mid \rho, \rho-a) = c_{n-k}(s \mid n-\rho, n-\rho-(k-a))$ and $C_k(s \mid \rho, a) = C_k(s \mid n-\rho, k-a) = C_{n-k}(s \mid \rho, \rho-a) = C_{n-k}(s \mid n-\rho, n-\rho-(k-a))$. Thus

$P(S_n \geq s \mid p, \rho, q) = P(S_n \geq s \mid q, n-\rho, p) = P(S_n \geq s \mid 1-p, \rho, 1-q)$
$= P(S_n \geq s \mid 1-q, n-\rho, 1-p)$.

(ii) For each n, k, j, ρ, a and $s_{\alpha,p}$, $d_k(j \mid \rho, a, s_{\alpha,p}) =$
$d_k(n-j \mid n-\rho, k-a, s_{\alpha,p}) = d_{n-k}(j \mid \rho, \rho-a, s_{\alpha,p}) = d_{n-k}(n-j \mid n-\rho, n-\rho-(k-a), s_{\alpha,p})$. Thus $P(\hat{\rho}_{n,\alpha} = j \mid p, \rho, q) = P(\hat{\rho}_{n,\alpha} = n-j \mid q, n-\rho, p) =$
$P(\hat{\rho}_{n,\alpha} = j \mid 1-p, \rho, 1-q) = P(\hat{\rho}_{n,\alpha} = n-j \mid 1-q, n-\rho, 1-q)$.

Again, as a check when counting note that $\sum_{a=0}^{\rho} c_k(s, j \mid \rho, a) =$
$c_k(s, j)$, $\sum_{a=0}^{\rho} c_k(s \mid \rho, a) = c_k(s)$ and $\sum_{a=0}^{\rho} d_k(j \mid \rho, a, s_{\alpha,p}) =$
$d_k(j \mid s_{\alpha,p})$ for each k, s, j, ρ, a and $s_{\alpha,p}$.

Properties of the test statistics and change point random variables under the alternative hypothesis are of interest when the null hypothesis is rejected, so expressions for the joint distributions need only be obtained for s^+ and s in the critical regions of the tests. Since $P(S_n^+ \geq 1 \mid p)$ and $P(S_n \geq 1 \mid p)$ are very large for each n and p, the critical regions contain only values $s^+ > 1$ and $s > 1$.

3.2 The Method

The method used to find expressions for the joint distributions uses the information obtained in III. of the procedure for counting in Section 2.2. Available from III. are the $c_k^+(s^+, j)$'s, $c_k(s, j)$'s and the types of sequences yielding each value of the test statistic. The manner in which the $c_k^+(s^+, j \mid \rho, a)$'s and $c_k(s, j \mid \rho, a)$'s are found

from this information is straightforward: for each ρ, $1 \leq \rho \leq n-1$, count the number of sequences which have a 1's in the first ρ places, $a \leq \rho$. The only means for reducing the counting task are the results in the previous section.

Application of the results in the previous section are illustrated by finding expressions for the joint distributions for $n = 5$.

Example 1. The $c_k^+(s^+, j \mid \rho, a)$'s are found for $n = 5$ and $s^+ = 2, \frac{3}{2}, \frac{4}{3}$. Expressions for the joint distributions are listed below.

$P\{(S_5^+, R_5^+) = (2, j) \mid p, \rho, q\}$		
$j = 1$	$\rho = 1$	$p(1-q)^4$
	$\rho = 2$	$p(1-p)(1-q)^3$
	$\rho = 3$	$p(1-p)^2(1-q)^2$
	$\rho = 4$	$p(1-p)^3(1-q)$
$j = 2$	$\rho = 1$	$pq(1-q)^3$
	$\rho = 2$	$p^2(1-q)^3$
	$\rho = 3$	$p^2(1-p)(1-q)^2$
	$\rho = 4$	$p^2(1-p)^2(1-q)$
$j = 3$	$\rho = 1$	$pq^2(1-q)^2$
	$\rho = 2$	$p^2q(1-q)^2$
	$\rho = 3$	$p^3(1-q)^2$
	$\rho = 4$	$p^3(1-p)(1-q)$

$j = 4$	$\rho = 1$	$pq^3(1-q)$
	$\rho = 2$	$p^2q^2(1-q)$
	$\rho = 3$	$p^3q(1-q)$
	$\rho = 4$	$p^4(1-q)$

$P\{(S_5^+, R_5^+) = (\frac{3}{2}, j) \mid p, \rho, q\}$

$j = 1$	$\rho = 1$	$3pq(1-q)^3$
	$\rho = 2$	$3p(1-p)q(1-q)^2$
	$\rho = 3$	$p^2(1-p)(1-q)^2 + 2p(1-p)^2q(1-q)$
	$\rho = 4$	$2p^2(1-p)^2(1-q) + p(1-p)^3q$
$j = 4$	$\rho = 1$	$2pq^2(1-q)^2 + (1-p)q^3(1-q)$
	$\rho = 2$	$p^2q(1-q)^2 + 2p(1-p)q^2(1-q)$
	$\rho = 3$	$3p^2(1-p)q(1-q)$
	$\rho = 4$	$3p^3(1-p)(1-q)$

$P\{S_5^+, R_5^+) = (\frac{4}{3}, j) \mid p, \rho, q\}$

$j = 2$	$\rho = 1$	$pq^2(1-q)^2$
	$\rho = 2$	$p^2q(1-q)^2$
	$\rho = 3$	$p^2(1-p)q(1-q)$
	$\rho = 4$	$p^2(1-p)^2q$
$j = 3$	$\rho = 1$	$(1-p)q^2(1-q)^2$
	$\rho = 2$	$p(1-p)q(1-q)^2$

$\rho = 3 \qquad p^2(1-p)(1-q)^2$

$\rho = 4 \qquad p^2(1-p)^2(1-q)$

When $(s^+, j) = (\frac{3}{2}, 4)$, for instance, the types of sequences involved are $T_5^3(4, 3) \setminus T_5^3(3, 3)$. There are three such sequences: -1111-1, 1-111-1, 11-11-1. For $\rho = 1$, two sequences have a 1 in the first place and one sequence has a -1 in the first place; $c_3^+(\frac{3}{2}, 4 \mid 1, 1) = 2$, $c_3^+(\frac{3}{2}, 4 \mid 1, 0) = 1$. For $\rho = 2$, one sequence has 2 1's in the first two places and two sequences have a 1 and a -1 in the first two places; $c_3^+(\frac{3}{2}, 4 \mid 2, 2) = 1$, $c_3^+(\frac{3}{2}, 4 \mid 2, 1) = 2$. For $\rho = 3$, all three sequences have 2 1's and 1 -1 in the first three places; $c_3^+(\frac{3}{2}, 4 \mid 3, 2) = 3$. For $\rho = 4$, all three sequences have 3 1's and 1 -1 in the first four places; $c_3^+(\frac{3}{2}, 4 \mid 4, 3) = 3$. Note that expressions for $P\{(S_5^+, R_5^+) = (\frac{3}{2}, 1) \mid p, \rho, q\}$ can be obtained from those for $P\{(S_5^+, R_5^+) = (\frac{3}{2}, 4) \mid p, \rho, q\}$ according to Theorem 1.

Example 2. The $c_k(s, j) \mid \rho, a)$'s are found for $n = 5$ and $s = 2, \frac{3}{2}, \frac{4}{3}$. Expressions for the joint distributions are listed below.

$P\{(S_5, R_5) = (2, j) \mid p, \rho, q\}$		
$j = 1$	$\rho = 1$	$p(1-q)^4 + (1-p)q^4$
	$\rho = 2$	$p(1-p)(1-q)^3 + p(1-p)q^3$
	$\rho = 3$	$p(1-p)^2(1-q)^2 + p^2(1-p)q^2$
	$\rho = 4$	$p(1-p)^3(1-q) + p^3(1-p)q$

$j = 2$	$\rho = 1$	$pq(1-q)^3 + (1-p)q^3(1-q)$
	$\rho = 2$	$p^2(1-q)^3 + (1-p)^2q^3$
	$\rho = 3$	$p^2(1-p)(1-q)^2 + p(1-p)^2q^2$
	$\rho = 4$	$p^2(1-p)^2(1-q) + p^2(1-p)^2q$
$j = 3$	$\rho = 1$	$pq^2(1-q)^2 + (1-p)q^2(1-q)^2$
	$\rho = 2$	$p^2q(1-q)^2 + (1-p)^2q^2(1-q)$
	$\rho = 3$	$p^3(1-q)^2 + (1-p)^3q^2$
	$\rho = 4$	$p^3(1-p)(1-q) + p(1-p)^2q$
$j = 4$	$\rho = 1$	$pq^3(1-q) + (1-p)q(1-q)^3$
	$\rho = 2$	$p^2q^2(1-q) + (1-p)^2q(1-q)^2$
	$\rho = 3$	$p^3q(1-q) + (1-p)^3q(1-q)$
	$\rho = 4$	$p^4(1-q) + (1-p)^4q$

$\underline{P\{(S_5, R_5) = (\frac{3}{2}, j) \mid p, \rho, q\}}$

$j = 1$	$\rho = 1$	$\frac{5}{2}pq(1-q)^3 + \frac{5}{2}(1-p)q^3(1-q)$
	$\rho = 2$	$\frac{5}{2}p(1-p)q(1-q)^2 + \frac{5}{2}p(1-p)q^2(1-q)$
	$\rho = 3$	$p^2(1-p)(1-q)^2 + \frac{3}{2}p(1-p)^2q(1-q) + p(1-p)^2q^2 + \frac{3}{2}p^2(1-p)q(1-q)$
	$\rho = 4$	$2p^2(1-p)^2(1-q) + \frac{1}{2}p(1-p)^3q + 2p^2(1-p)^2q + \frac{1}{2}p^3(1-p)(1-q)$

j = 4 $\quad \rho = 1 \quad 2pq^2(1-q)^2 + \frac{1}{2}(1-p)q^3(1-q) + 2(1-p)q^2(1-q)^2 + \frac{1}{2}pq(1-q)^3$

$\rho = 2 \quad p^2q(1-q)^2 + \frac{3}{2}p(1-p)q^2(1-q) + (1-p)^2q^2(1-q) + \frac{3}{2}p(1-p)q(1-q)^2$

$\rho = 3 \quad \frac{5}{2}p^2(1-p)q(1-q) + \frac{5}{2}p(1-p)^2q(1-q)$

$\rho = 4 \quad \frac{5}{2}p^3(1-p)(1-q) + \frac{5}{2}p(1-p)^3q$

$\underline{P\{(S_5, R_5) = (\frac{4}{3}, j) \mid p, \rho, q\}}$

j = 2 $\quad \rho = 1 \quad pq^2(1-q)^2 + (1-p)q^2(1-q)^2$

$\rho = 2 \quad p^2q(1-q)^2 + (1-p)^2q^2(1-q)$

$\rho = 3 \quad p^2(1-p)q(1-q) + p(1-p)^2q(1-q)$

$\rho = 4 \quad p^2(1-p)^2q + p^2(1-p)^2(1-q)$

j = 3 $\quad \rho = 1 \quad (1-p)q^2(1-q)^2 + pq^2(1-q)^2$

$\rho = 2 \quad p(1-p)q(1-q)^2 + p(1-p)q^2(1-q)$

$\rho = 3 \quad p^2(1-p)(1-q)^2 + p(1-p)^2q^2$

$\rho = 4 \quad p^2(1-p)^2(1-q) + p^2(1-p)^2q$

Consider the pair $(s, j) = (\frac{3}{2}, 4)$; the $c_k(\frac{3}{2}, 4 \mid \rho, a)$'s need only be counted for $k \geq \left[\frac{5}{2}\right]$. The three sequences involved are -1111-1, 1-111-1, 11-11-1; the first sequence has $J_5 = 1, 4$ so it contributes $\frac{1}{2}$ to the count. Note that $c_3(\frac{3}{2}, 4 \mid 1, 1) = 2$, $c_3(\frac{3}{2}, 4 \mid 1, 0) = \frac{1}{2}$,

$c_3(\frac{3}{4}, 4 \mid 2, 2) = 1$, $c_3(\frac{3}{2}, 4 \mid 2, 1) = \frac{3}{2}$, $c_3(\frac{3}{2}, 4 \mid 3, 2) = \frac{5}{2}$ and $c_3(\frac{3}{2}, 4 \mid 4, 3) = \frac{5}{2}$. The $c_2(\frac{3}{2}, 4 \mid \rho, a)$'s , $c_2(\frac{3}{2}, 1 \mid \rho, a)$'s and $c_3(\frac{3}{2}, 1 \mid \rho, a)$'s can be found from the $c_3(\frac{3}{2}, 4 \mid \rho, a)$'s according to Theorem 3.

The marginal distributions S_5^+, $\hat{\rho}_{5,\alpha}^+$, S_5 and $\hat{\rho}_{5,\alpha}$ may be obtained from the joint distributions.

3.3 The One-Sided Testing Situation

The behavior of the joint densities (S_n^+, R_n^+) is analyzed; the resulting behavior of the power function and estimator for the change point are discussed. Lastly, properties of $\hat{q}$ are investigated.

3.3.1 The Joint Densities

The following result describes the local behavior of the joint density.

Theorem 5. For each n, s^+, j, p and q where $p > q$,

$$P\{(S_n^+, R_n^+) = (s^+, j) \mid p, \rho=j, q\} \geq P\{(S_n^+, R_n^+) = (s^+, j) \mid p, \rho=j\pm 1, q\}.$$

Proof. Suppose $x_1, \cdots, x_n$ has $S_n^+ = s^+$ and $R_n^+ = j$; then $P(x_1, \cdots, x_n \mid p, \rho=j, q) = p^a(1-p)^{j-a}q^{k-a}(1-q)^{n-j-(k-a)}$ if $x_1, \cdots, x_n$ has k 1's. Also $P(x_1, \cdots, x_n \mid p, \rho=j-1, q) = p^{a-1}(1-p)^{j-a}q^{k+1-a} \cdot (1-q)^{n-j-(k-a)}$ and $P(x_1, \cdots, x_n \mid p, \rho=j+1, q) = p^a(1-p)^{j+1-a} \cdot q^{k-a}(1-q)^{n-1-j-(k-a)}$ since $x_j = 1$ and $x_{j+1} = -1$.

Unfortunately, Theorem 5 cannot be generalized to state that $P\{(S_n^+, R_n^+) = (s^+, j) \mid p, \rho = j\pm i, q\}$ $P\{(S_n^+, R_n^+) = (s^+, j) \mid p, \rho = j\pm(i+1), q\}$; note that $P\{(S_5^+, R_5^+) = (\frac{3}{2}, 1) \mid p, \rho=2, q\} = 3p(1-p)q(1-q)^2 \geq p^2(1-p)(1-q)^2 + 2p(1-p)^2 q(1-q) = P\{(S_5^+, R_5^+) = (\frac{3}{2}, 1) \mid p, \rho=3, q\}$ if and only if $q \geq \frac{1}{3}$.

The following properties describe general behaviors of the joint density; computation indicates that each is true but derivation for arbitrary n has not been possible.

<u>Property 1.</u> For each n, s^+, $i > 0$, p and q where $p > q$,

$$P\{(S_n^+, R_n^+) = (s^+, j) \mid p, \rho=j+1, q\} \geq P\{(S_n^+, R_n^+) = (s^+, j) \mid p, \rho=j+i+1, q\}$$

when

(i) $j < \left[\frac{n}{2}\right]$ and $\frac{1}{2} \leq q < p$

and (ii) $j > \left[\frac{n}{2}\right]$;

$$P\{(S_n^+, R_n^+) = (s^+, j) \mid p, \rho=j-i, q\} \geq P\{(S_n^+, R_n^+) = (s^+, j) \mid p, \rho=j-(i+1), q\}$$

when

(i) $j < \left[\frac{n}{2}\right]$

and (ii) $j > \left[\frac{n}{2}\right]$ and $q < p \leq \frac{1}{2}$.

For $j < \left[\frac{n}{2}\right]$ the first inequality in Property 1 may be reversed for some i when $p > \frac{1}{2} > q$ and q is small or when $\frac{1}{2} > p > q$ and q is small; for $j > \left[\frac{n}{2}\right]$ the second inequality may be reversed for some i if p is large. For $j = \left[\frac{n}{2}\right]$ the inequalities may be reversed for some i when p is large and q is small.

Summing the probability inequalities over values $s^+ \geq s^+_{\alpha,p} > 1$ yields global properties.

Property 2. For each n, $s^+_{\alpha,p} > 1$, j, i > 0, p and q where p > q,

$$\sum_{s^+ \geq s^+_{\alpha,p}} P\{(S^+_n, R^+_n) = (s^+, j) \mid p, \rho = j+i, q\}$$
$$> \sum_{s^+ \geq s^+_{\alpha,p}} P\{(S^+_n, R^+_n) = (s^+, j) \mid p, \rho = j+i+1, q\}$$

and
$$\sum_{s^+ \geq s^+_{\alpha,p}} P\{(S^+_n, R^+_n) = (s^+, j) \mid p, \rho = j-i, q\}$$
$$> \sum_{s^+ \geq s^+_{\alpha,p}} P\{(S^+_n, R^+_n) = (s^+, j) \mid p, \rho = j-(i+1), q\}.$$

That is, the probability that $S^+_n \geq s^+_{\alpha,p}$ and $R^+_n = j$ is largest when $\rho = j$ and the probability decreases monotonically as ρ increases or decreases from j.

Property 3. For each n, $s^+_{\alpha,p} > 1$, j, p and q where p > q,

$$\sum_{s^+ \geq s^+_{\alpha,p}} P\{(S^+_n, R^+_n) = (s^+, j) \mid p, \rho = j+i, q\}$$
$$\geq \sum_{s^+ \geq s^+_{\alpha,p}} P\{(S^+_n, R^+_n) = (s^+, j) \mid p, \rho = j-i, q\}$$

when $i > 0$ and $p \leq \frac{1}{2}$, or $i > 0$ and $q \leq 1-p < p$, or $i < 0$ and $1-p \leq q < p$.

<u>Property 4.</u> Although no monotonic properties hold for the joint probabilities when $R_n^+ = j_1$ and $R_n^+ = j_2$ the following general tendency is observed:

$$\sum_{s^+ \geq s^+_{\alpha,p}} P\{(S_n^+, R_n^+) = (s^+, j_1) \mid p, \rho = j_1, q\}$$

$$\geq \sum_{s^+ \geq s^+_{\alpha,p}} P\{(S_n^+, R_n^+) = (s^+, j_2) \mid p, \rho = j_2, q\}$$

when $j_1 < j_2$, $p \leq \frac{1}{2}$ or $j_1 < j_2$, $q < 1-p < p$ or $j_2 < j_1$, $1-p < q < p$.

For fixed ρ and varying q the joint probabilities behave according to Property 5.

<u>Property 5.</u> For each n, $s^+_{\alpha,p} > 1$, j, $i \geq 0$, p, q_1 and q_2 where $p > q_1 > q_2$,

$$\sum_{s^+ \geq s^+_{\alpha,p}} P\{(S_n^+, R_n^+) = (s^+, j) \mid p, \rho = j+i, q_1\}$$

$$\leq \sum_{s^+ \geq s^+_{\alpha,p}} P\{(S_n^+, R_n^+) = (s^+, j) \mid p, \rho = j+i, q_2\} .$$

Thus the probability that $S_n^+ \geq s^+_{\alpha,p}$ and $R_n^+ = j$ increases for fixed $\rho > j$ and p as q decreases. For $i < 0$ the inequality in Property 5 may be reversed; in general the inequality is true for a larger set of $i < 0$ and q_2 when j is small and/or p is large.

3.3.2 The Power Function

The power function $P(S_n^+ \geq s_{\alpha,p}^+ | p, \rho, q)$ is analyzed as a function of ρ and q. Power functions for $n = 10, 20$, $\alpha \doteq .06$ and $p = .3, .5$, $\alpha \doteq .10$ and $p = .7, .9$, and various values of ρ are found in Figures 30.1-30.4; the figures illustrate the following typical behavior of the power functions for $n \leq 20$. For $p \leq \frac{1}{2}$ the power increases as q decreases from p for each ρ and the power tends to increase as ρ decreases for each q; see Properties 4, 5. For small n one can show algebraically that the power increases as q decreases for each ρ, but the manipulations required to verify this for larger n are horrendous. Also for $p \leq \frac{1}{2}$ the power tends to increase as n increases for each q and $\rho \doteq [cn]$, $c \leq 1$.

For $p > \frac{1}{2}$ the test is slightly biased; the bias is occurring for q near $\frac{1}{2}$ and ρ small, as is explained by Properties 3, 4. As α increases the bias disappears. A further explanation for the bias is provided by the properties (A1) and (A2) of $P(S_n^+ \geq s^+ | p)$ in Section 2.3; if the tendency for the maximums to occur at $p = \frac{1}{2}$ for smaller s^+ values as n increases is true, one expects that the power function will remain biased for q near $\frac{1}{2}$ and ρ small as n increases. The power is not necessarily an increasing function in n for $p > \frac{1}{2}$ unless ρ and q are sufficiently small.

The power of S_n^+, the generalized likelihood ratio statistic $C_{p,q}$ and Page's test statistic M are tested in Table 31 for $n = 5$ and $p = .4$,

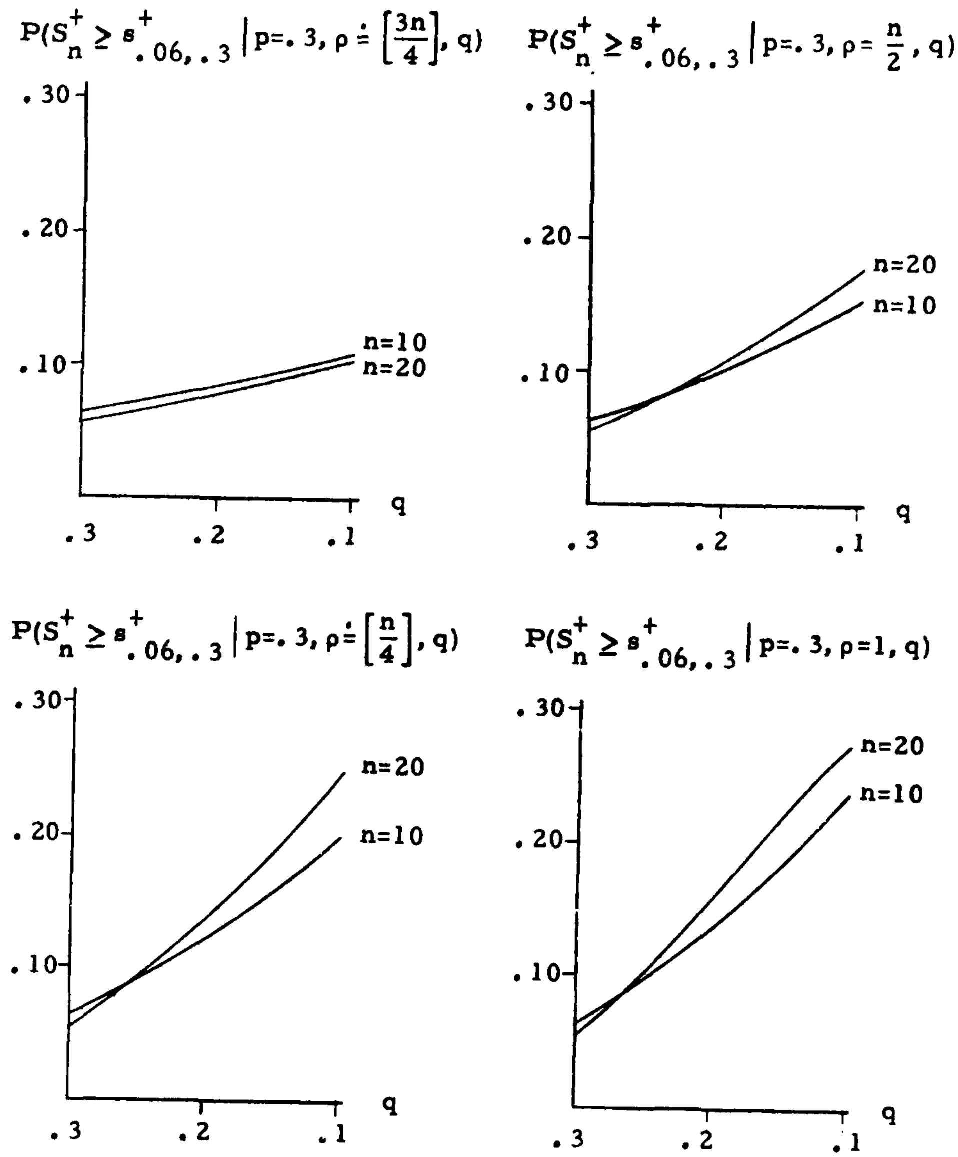


Figure 30.1. $P(S_n^+ \geq s^+_{.06,.3} \mid p=.3, \rho, q)$.

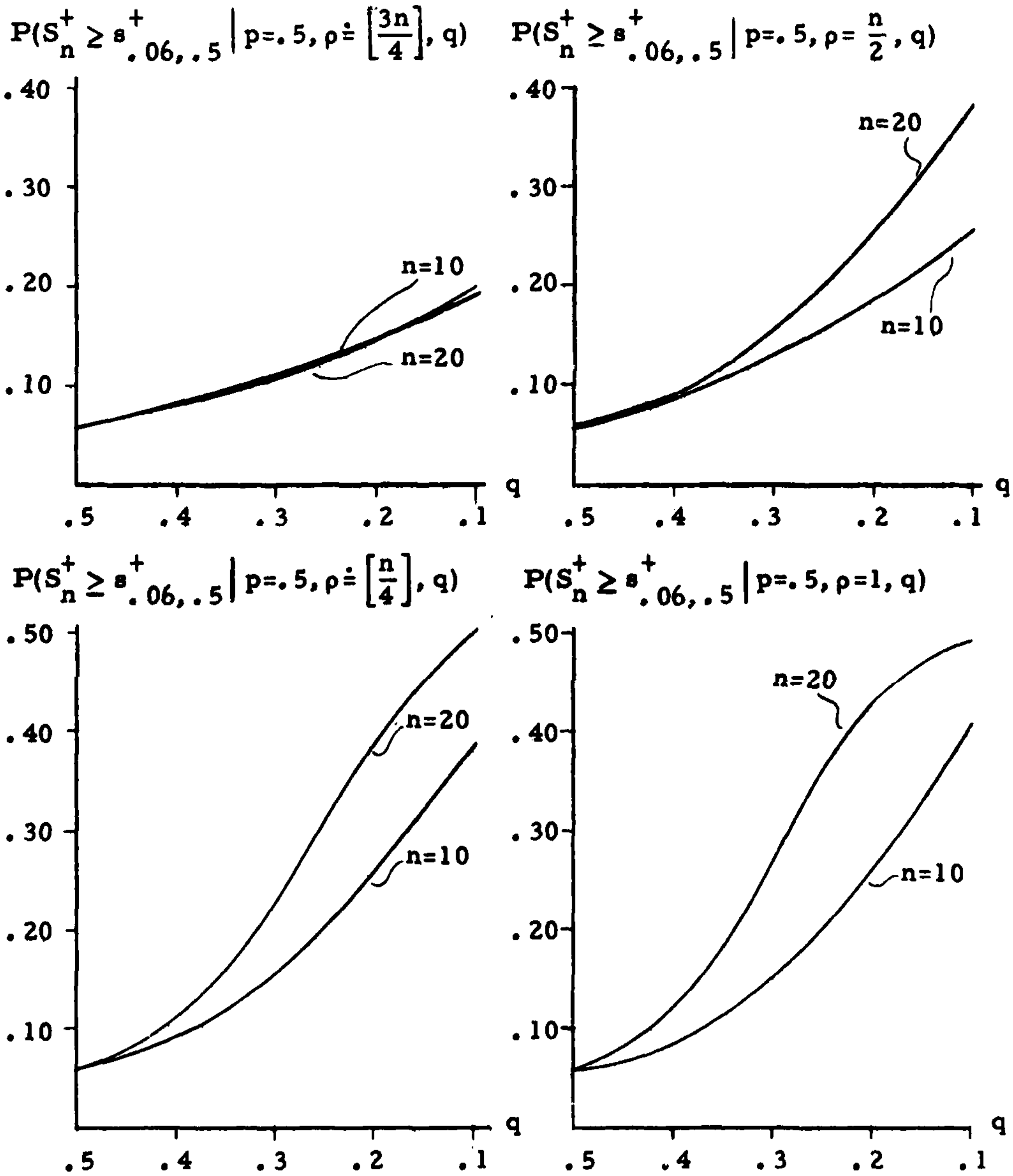


Figure 30.2. $P(S_n^+ \geq s^+_{.06,.5} \mid p=.5, \rho, q)$.

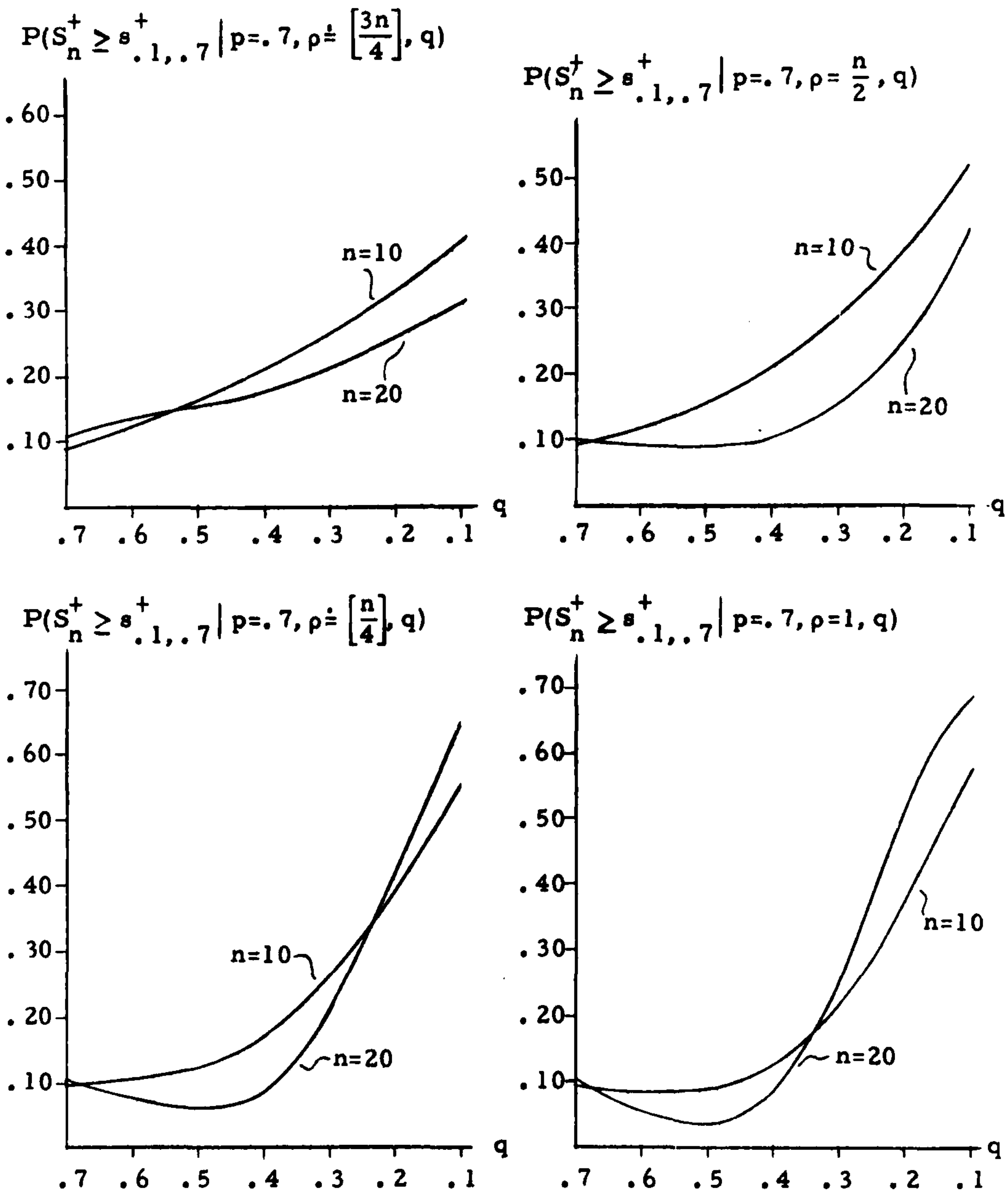


Figure 30.3. $P(S_n^+ \geq s^+_{.1,.7} \mid p=.7, \rho, q)$.

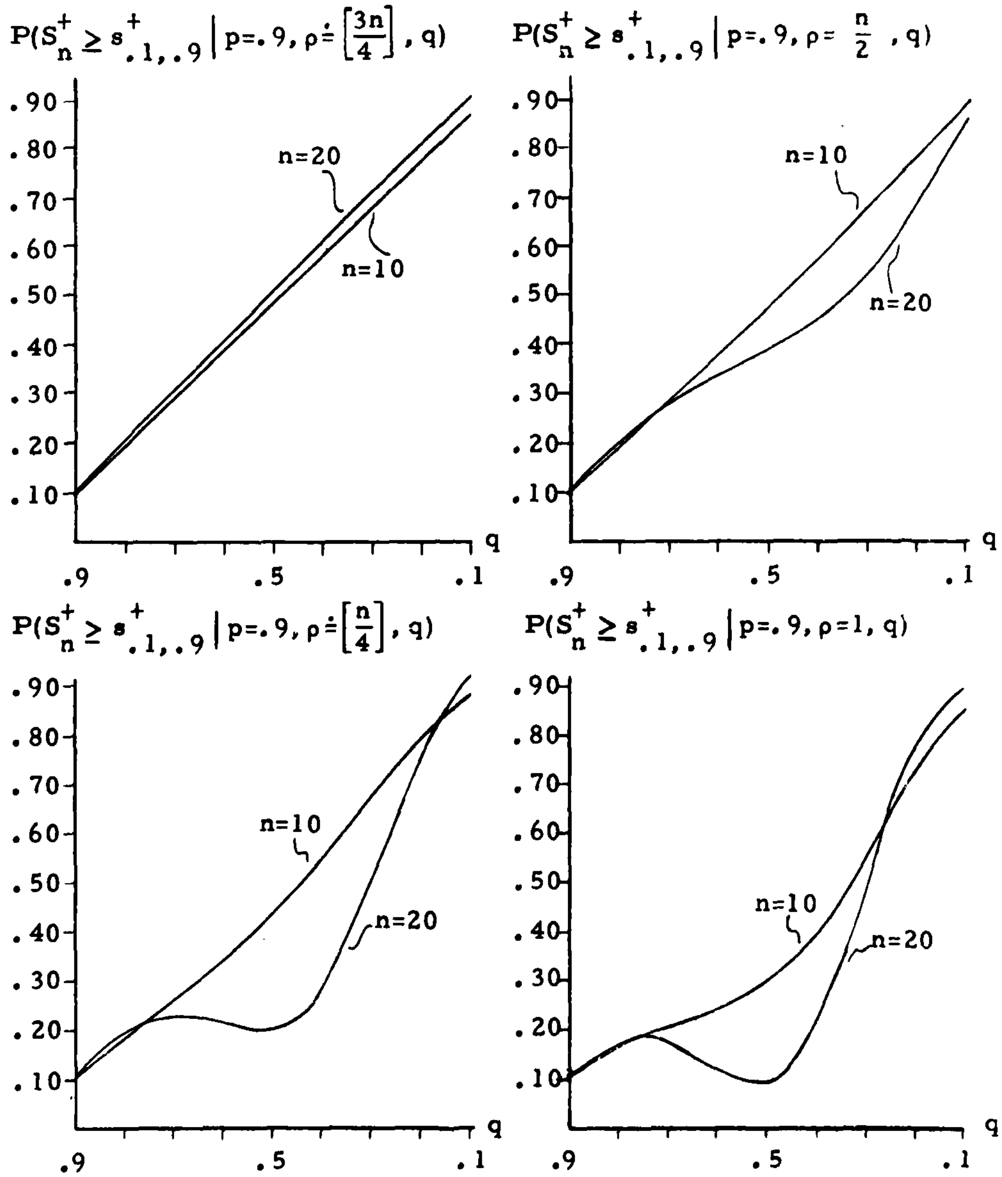


Figure 30.4. $P(S_n^+ \geq s^+_{.1,.9} \mid p = .9,\ \rho,\ q)$.

Table 31. The power of S_n^+, $C_{p,q}$ and M for a sample of size n=5.

p=.4					
S_5^+	4	.193	.172	.151	.130
$C_{.4,q}$	4	.194	.173	.151	.130
M	4	.162	.151	.140	.130
S_5^+	3	.256	.209	.167	.130
$C_{.4,q}$	3	.292	.230	.176	.130
M	3	.243	.202	.164	.130
S_5^+	2	.308	.234	.175	.130
$C_{.4,q}$	2	.437	.307	.206	.130
M	2	.365	.269	.191	.130
S_5^+	1	.298	.222	.167	.130
$C_{.4,q}$	1	.656	.410	.240	.130
M	1	.547	.358	.223	.130
	p/q	.1	.2	.3	.4

Table 31. (Continued).

p=.5						
S_5^+	4	.169	.150	.132	.113	.094
$C_{.5,q}$	4	.169	.150	.131	.113	.094
M	4	.119	.113	.106	.100	.094
S_5^+	3	.237	.196	.158	.124	.094
$C_{.5,q}$	3	.304	.240	.184	.135	.094
M	3	.214	.180	.149	.120	.094
S_5^+	2	.291	.223	.168	.126	.094
$C_{.5,q}$	2	.547	.384	.257	.162	.094
M	2	.385	.288	.208	.144	.094
S_5^+	1	.277	.205	.153	.117	.094
$C_{.5,q}$	1	.693	.461	.292	.173	.094
M	1	.693	.461	.292	.173	.094
	ρ/q	.1	.2	.3	.4	.5

Table 31. (Continued).

p=.7					
S_5^+	4	.365	.284	.203	.122
$C_{.7,q}$	4	.320	.250	.185	.122
M	4	.203	.176	.150	.123
S_5^+	3	.479	.343	.224	.122
$C_{.7,q}$	3	.824	.504	.278	.122
M	3	.441	.317	.211	.123
S_5^+	2	.554	.343	.210	.122
$C_{.7,q}$	2	.851	.564	.313	.122
M	2	.821	.518	.286	.123
S_5^+	1	.517	.284	.175	.122
$C_{.7,q}$	1	.933	.677	.366	.122
M	1	.955	.691	.369	.123
	ρ/q	.1	.3	.5	.7

Table 31. (Continued).

p=.9						
S_5^+	4	.664	.517	.369	.221	.074
$C_{.9,q}$	4	.365	.446	.217	.130	.074
M	4	.314	.254	.194	.134	.074
S_5^+	3	.729	.554	.387	.227	.074
$C_{.9,q}$	3	.844	.637	.367	.206	.074
M	3	.852	.589	.372	.201	.074
S_5^+	2	.729	.479	.315	.194	.074
$C_{.9,q}$	2	.878	.820	.368	.292	.074
M	2	.976	.815	.561	.290	.074
S_5^+	1	.664	.365	.225	.157	.074
$C_{.9,q}$	1	.965	.829	.593	.356	.074
M	1	.997	.922	.706	.379	.074
	p/q	.1	.3	.5	.7	.9

.5, .7, .9. To attain approximately the same significance level for each p it was necessary to randomize; the procedure used was to assign the required fraction of probability to each sequence for which the value of the statistic overlapped the critical region.

The powers of the tests were not computed for $p < .4$ since for the generalized likelihood ratio test and Page's test the null hypothesis cannot be rejected at a reasonable α level for small p. For the test based on S_5^+ one can reject H_o for the smaller values of p, however, for reasonable α levels since $P(S_5^+ \geq s \mid p) = P(S_5^+ \geq s \mid 1-p)$ for all p.

Table 31 shows that S_5^+ has greater power than M when the change occurs later in the sequence; M has greater power otherwise. Hence it seems that the test based on S_n^+ is better suited for detecting a change later in the sequence than is M. Recall the difficulty with Page's statistic in estimating change points later in the sequence; also recall that the types of sequences in the critical region of his test are not in general indicative of sequences for which a change in the probability of a 1 has decreased: $x_1, \cdots, x_n$ and $x_n, \cdots, x_1$ give the same value of the statistic. For $p \geq \frac{1}{2}$ it appears as though M is constructed to behave more like a likelihood ratio test than a change point test.

The power of the generalized likelihood ratio test is larger than that of Page's test unless $p \geq \frac{1}{2}$ and ρ is small.

The power of the test based on S_5^+ is larger than the power of the test based on $C_{p,q}$ when ρ is large; otherwise the opposite is true. Hence here also it seems that the test based on S_n^+ is better suited for detecting a change later in the sequence than is $C_{p,q}$. Moreover $C_{p,q}$ suffers from the same drawback as M in not being able to estimate change points later in the sequence; $C_{p,q}$ and M both worsen in this respect as p decreases. It is incredibly more difficult to evaluate $C_{p,q}$ for a sample sequence and to obtain critical regions and properties of the test than with either S_n^+ or M ; also note that although q is in general unknown the critical region of the test based on $C_{p,q}$ depends on q as well as an p .

Table 32 lists the power of Page's statistic for $n = 20$, $p = \frac{1}{2}$ and $\alpha = .05$ as computed by Chernoff and Zachs (1964) and the power of S_{20}^+ for $p = \frac{1}{2}$, $\alpha = .057$. Here also S_{20}^+ has greater power for larger values of ρ and M has greater power otherwise.

3.3.3 The Estimator of the Change Point

Property 2 indicates that the mode of the probability that $S_n^+ \geq s_{\alpha,p}^+$ and $R_n^+ = j$ occurs at $\rho = j$ and that the probability decreases monotonically as ρ increases or decreases from j; Property 3 further describes the manner in which the probabilities decrease as ρ increases or decreases from j. These properties

Table 32. The power of S_n^+ and Page's test statistic M for a sample of size n = 20, $\alpha = .05$.

S_{20}^+	18	.098	.084	.071	.057
M	18	.066	.060	.055	.050
S_{20}^+	16	.125	.097	.075	.057
M	16	.105	.082	.064	.050
S_{20}^+	14	.177	.120	.079	.057
M	14	.171	.116	.076	.050
S_{20}^+	12	.207	.132	.084	.057
M	12	.272	.162	.092	.050
S_{20}^+	10	.259	.156	.090	.057
M	10	.407	.223	.111	.050
S_{20}^+	8	.316	.185	.097	.057
M	8	.545	.292	.132	.050
S_{20}^+	6	.365	.215	.104	.057
M	6	.664	.364	.154	.050
S_{20}^+	4	.410	.249	.113	.057
M	4	.756	.431	.176	.050
S_{20}^+	2	.437	.276	.120	.057
M	2	.821	.488	.195	.050
	p/q	.2	.3	.4	.5

and the behavior of the power function are useful for analyzing $P(\hat{\rho}^{+}_{n,\alpha} = j \mid p, \rho, q)$ as a function of its parameters.

For each $n \leq 20$ computation indicates that $P(\hat{\rho}^{+}_{n,\alpha} = j \mid p, \rho, q)$ also has a mode at $\rho = j$; moreover as ρ increases or decreases from j, the probability generally monotonically decreases. The manner in which $P(\hat{\rho}^{+}_{n,\alpha} = j \mid p, \rho = j+1, q)$ decreases for $i > 0$ or $i < 0$ is as follows: $P(\hat{\rho}^{+}_{n,\alpha} = j \mid p, \rho = j+i, q) > P(\hat{\rho}^{+}_{n,\alpha} = j \mid p, \rho = j-i, q)$ when $p \leq \frac{1}{2}$, $i > 0$ or $q < 1-p < p$, $i > 0$ or $1-p < q < p$, $i < 0$.

As a function in q, $P(\hat{\rho}^{+}_{n,\alpha} = j \mid p, \rho = j+i, q)$ in general decreases as $q \to p$ for $i \geq 0$, increases as $q \to p \leq \frac{1}{2}$, for $i < 0$, and may increase and then decrease as $q \to p > \frac{1}{2}$ for $i < 0$.

The magnitude of the modal probability $P(\hat{\rho}^{+}_{n,\alpha} = j \mid p, \rho = j, q)$ relative to $\sum_i P(\hat{\rho}^{+}_{n,\alpha} = j \mid p, \rho = j \pm i, q)$ depends heavily on j, p and q. For each j, as p-q increases the relative magnitude of the model probability increases; for $p \leq \frac{1}{2}$ the relative magnitude for $j > \left[\frac{n}{2}\right]$ tends to be greater than that for $j < \left[\frac{n}{2}\right]$ for each q; for $p > \frac{1}{2}$ the relative magnitude for $j < \left[\frac{n}{2}\right]$ tends to be larger than for $j > \left[\frac{n}{2}\right]$ when $1-p < q < p$ and smaller when $q < 1-p < p$.

Table 33 illustrates the behavior of $P(\hat{\rho}^{+}_{n,\alpha} = j \mid p, \rho, q)$ for $n=10$, $j=1,5,9$ and $p = .5, .9$. Observe that the mode of the probabilities occurs at $\rho = j$ for each α, j, p and q. Also observe the probabilities decrease quickly for $\rho < j$, $p = .5$ and $\rho > j$, $p = .9$ whereas the probabilities decrease much more slowly for $\rho > j$,

Table 33. $P(\hat{\rho}^{+}_{10,\alpha} = j \mid p, \rho, q)$.

$p = .5, j = 1, \alpha = .050$

ρ/q	.1	.3
9	.1678	.1710
8	.1917	.1852
7	.2337	.2131
6	.3139	.2605
5	.3504	.2972
4	.4003	.3337
3	.4373	.3591
2	.4885	.4190
1	.8958	.6260

$p = .5, j = 5, \alpha = .050$

ρ/q	.1	.3
9	.0207	.0203
8	.0265	.0235
7	.0369	.0293
6	.0577	.0388
5	.0769	.0484
4	.0121	.0257
3	.0019	.0132
2	.0004	.0074
1	.0001	.0048

$p = .5, j = 9, \alpha = .050$

ρ/q	.1	.3
9	.1862	.1825
8	.0265	.0909
7	.0078	.0556
6	.0020	.0357
5	.0004	.0212
5	.0001	.0124
3	.0000	.0070
2	.0000	.0043
1	.0000	.0029

$p = .9, j = 1, \alpha = .044$

ρ/q	.1	.5	.7
9	.0000	.0000	.0000
8	.0000	.0000	.0000
7	.0000	.0000	.0000
6	.0000	.0000	.0000
5	.0001	.0000	.0000
4	.0011	.0002	.0000
3	.0099	.0017	.0000
2	.0899	.0136	.0003
1	.8389	.1113	.0007

$p = .9, j = 5, \alpha = .044$

ρ/q	.1	.5	.7
9	.0001	.0001	.0002
8	.0011	.0007	.0004
7	.0099	.0040	.0014
6	.0889	.0269	.0052
5	.8000	.1951	.0194
4	.0889	.1632	.0194
3	.0099	.1405	.0194
2	.0011	.1231	.0194
1	.0001	.1113	.0195

$p = .9, j = 9, \alpha = .044$

ρ/q	.1	.5	.7
9	.8889	.8889	.8889
8	.0899	.4706	.6747
7	.0099	.3200	.6116
6	.0011	.2425	.5880
5	.0001	.1951	.5783
4	.0000	.1632	.5744
3	.0000	.1405	.5727
2	.0000	.1231	.5719
1	.0000	.1113	.5693

$p = .5$ and $\rho < j$, $p = .9$. Thus for example note that

$$\frac{P(\hat{\rho}^{+}_{10,.050} = 1 \mid p=.5,\ \rho=1,\ q=.1)}{\sum_i P(\hat{\rho}^{+}_{10,.050} = 1 \mid p=.5,\ \rho=1+i,\ q=.1)} = \frac{.8958}{3.4794} = .2575$$

whereas

$$\frac{P(\hat{\rho}^{+}_{10,.050} = 9 \mid p=.5,\ \rho=9,\ q=.1)}{\sum_i P(\hat{\rho}^{+}_{10,.050} = 9 \mid p=.5,\ \rho=9-i,\ q=.1)} = \frac{.1862}{.2230} = .8350 .$$

Since the magnitude of the modal probability to $\sum_i P(\hat{\rho}^{+}_{n,\alpha} = j \mid p,\ \rho=j\pm i, q)$ is a measure of the goodness of the estimator and since in application it is desirable to detect change points quickly, the following experimental design appears to be optimal. Dichotomize the observations $Y_1, \cdots, Y_n$ to obtain the observations

$$X_i = \begin{cases} 1 & \text{if} \quad Y_i \geq g(\theta) \\ -1 & \text{if} \quad Y_i < g(\theta) \end{cases}$$

where θ is the location parameter under the null hypothesis and g is defined so that $P(Y_i \geq g(\theta) \mid \theta) \leq \frac{1}{2}$; thus $P(X_i = 1 \mid p = \frac{1}{2}) \leq \frac{1}{2}$. Furthermore sample the Y_i's if possible so that an anticipated change point occurs late in the sequence.

3.3.4 The Estimator of q

When the estimated change point $\hat{\rho}$ is greater than or equal to the actual change point ρ , the last $n-\hat{\rho}$ observations can be used to

obtain an unbiased estimator $\hat{q} = \frac{1}{n-\hat{\rho}} \sum_{i=\hat{\rho}+1} \frac{x_{i+1}}{2}$ of q. If $\hat{\rho} < \rho$ the last $n-\hat{\rho}$ observations give a biased estimator for q however.

3.4 The Two-Sided Testing Situation

In the last section it was noted that the one-sided test is unbiased for $p \leq \frac{1}{2}$, but that for $p > \frac{1}{2}$ the test is biased for ρ small and q near $\frac{1}{2}$, and the bias increases as p increases from $\frac{1}{2}$. The structure producing the bias for the one-sided test is present also for the two-sided test where under the alternative hypothesis q may be greater or less than p ; this test is unbiased for $p = \frac{1}{2}$, biased for $p \neq \frac{1}{2}$ and the bias increases as p deviates from $\frac{1}{2}$. The behavior of the joint densities (S_n, R_n) are studied for $p = \frac{1}{2}$; the resulting behavior of the power function and estimator for the change point are investigated. Lastly, properties of $\hat{q}$ are discussed.

3.4.1 The Joint Densities

The following properties describe the general behavior of the joint density for $p = \frac{1}{2}$. Property 6 describes the local behavior.

Property 6. For each n, s, j and q, $P\{(S_n, R_n) = (s,j) \mid p = \frac{1}{2}, \rho=j, q\} \geq P\{(S_n, R_n) = (s,j) \mid p = \frac{1}{2}, \rho=j\pm1, q\}$.

Summing the probability inequalities over values $s \geq s_\alpha$ yields the following global properties.

<u>Property 7.</u> For each n, $s_\alpha > 1$, $i \geq 0$ and q,

$$\sum_{s \geq s_\alpha} P\{(S_n, R_n) = (s, j) \mid p = \tfrac{1}{2}, \ \rho = j+i, \ q\}$$

$$\geq \sum_{s \geq s_\alpha} P\{(S_n, R_n) = (s, j) \mid p = \tfrac{1}{2}, \ \rho = j+i+1, \ q\}$$

for all j, and $\sum_{s \geq s_\alpha} P\{(S_n, R_n) = (s, j) \mid p = \tfrac{1}{2}, \ j-i, \ q\}$

$$\geq \sum_{s \geq s_\alpha} P\{(S_n, R_n) = (s, j) \mid p = \tfrac{1}{2}, \ \rho = j-i-1, q\}$$

for $j \leq \left[\frac{n}{2}\right]$.

Thus when $j \leq \left[\frac{n}{2}\right]$, the probability that $S_n \geq s_\alpha$ and $R_n = j$ is largest when $\rho = j$ and the probability decreases monotonically as ρ increases or decreases from j; when $j > \left[\frac{n}{2}\right]$ and $\rho < j$ this may not be true.

The following property describes the manner in which the probabilities decrease as ρ deviates from j.

Property 8. For each n, $s_\alpha > 1$, j, $i \geq 0$ and q,

$$\sum_{s \geq s_\alpha} P\{(S_n, R_n) = s, j) \,|\, p = \frac{1}{2}, \ \rho = j+i, q\}$$

$$\sum_{s \geq s_\alpha} P\{(S_n, R_n) = (s, j) \,|\, p = \frac{1}{2}, \ \rho = j-i, \ q\} \ .$$

Property 9. For each n, $s_\alpha > 1$, q and $j < \left[\frac{n}{2}\right]$, the tendency is that

$$\sum_{s \geq s_\alpha} P\{(S_n, R_n) = (s, j) \,|\, p = \frac{1}{2}, \ \rho = j, q\}$$

$$\sum_{s \geq s_\alpha} P\{(S_n, R_n) = (s, n-j) \,|\, p = \frac{1}{2}, \ \rho = n-j, \ q\} \ .$$

For fixed ρ and varying q the joint probabilities behave according to Property 10.

Property 10. For each n, $s_\alpha > 1$, j, $i \geq 0$, q_1 and q_2 where $q_1 < q_2 < p$ or $q_1 > q_2 > p$,

$$\sum_{s \geq s_\alpha} P\{(S_n, R_n) = (s, j) \,|\, p = \frac{1}{2}, \ \rho = j+i, \ q_1\}$$

$$\sum_{s \geq s_\alpha} P\{(S_n, R_n) = (s, j) \,|\, p = \frac{1}{2}, \ \rho = j+i, \ q_2\} \ .$$

3.4.2 The Power Function

The power function $P(S_n \geq s_{\alpha,p} | p, \rho, q)$ is studied as a function of ρ and q. Power functions for $n = 10, 20$ and $p = .5$, $.3$ and various values of ρ are found in Figures 34.1, 34.2. For $p = .5$, $\alpha = .047$, $.049$ for $n = 10, 20$, respectively; for $p = .3$, $\alpha = .121$, $.095$ for $n = 10, 20$, respectively. The figures illustrate the following typical behavior of the power function for $n \leq 20$. For $p = .5$ the test is unbiased and the power tends to increase as $|q-p|$ increases for each ρ and as ρ decreases for q. Also the power tends to increase as n increases for each q and $\rho \doteq [cn]$, $c \leq 1$. For $p = .3$ the figure illustrates the bias of the test as q increases from p; the test is unbiased for small and large alternatives q.

3.4.3 The Estimator of the Change Point

Properties 7, 8 and 10 describe the behavior of the probability that $S_n \geq s_\alpha$ and $R_n = j$ as functions of ρ and q. Computations indicate that the same behaviors occur for $\hat{\rho}_{n,\alpha}$, $n \leq 20$. That is, the mode of $P(\hat{\rho}_{n,\alpha} = j | p = \frac{1}{2}, \rho, q)$ occurs at $\rho = j$ when $j \geq \left[\frac{n}{2}\right]$ and the probabilities decrease monotonically as ρ increases or decreases from j; when $j > \left[\frac{n}{2}\right]$ the probabilities do not monotonically decrease as ρ decreases from j. The manner in which the probabilities behave for $\rho = j+i$, $j-i$ is as follows: for $i > 0$, $P(\hat{\rho}_{n,\alpha} = j | p = \frac{1}{2}, \rho = j+i, q) > P(\hat{\rho}_{n,\alpha} = j | p = \frac{1}{2}, \rho = j-i, q)$. As a

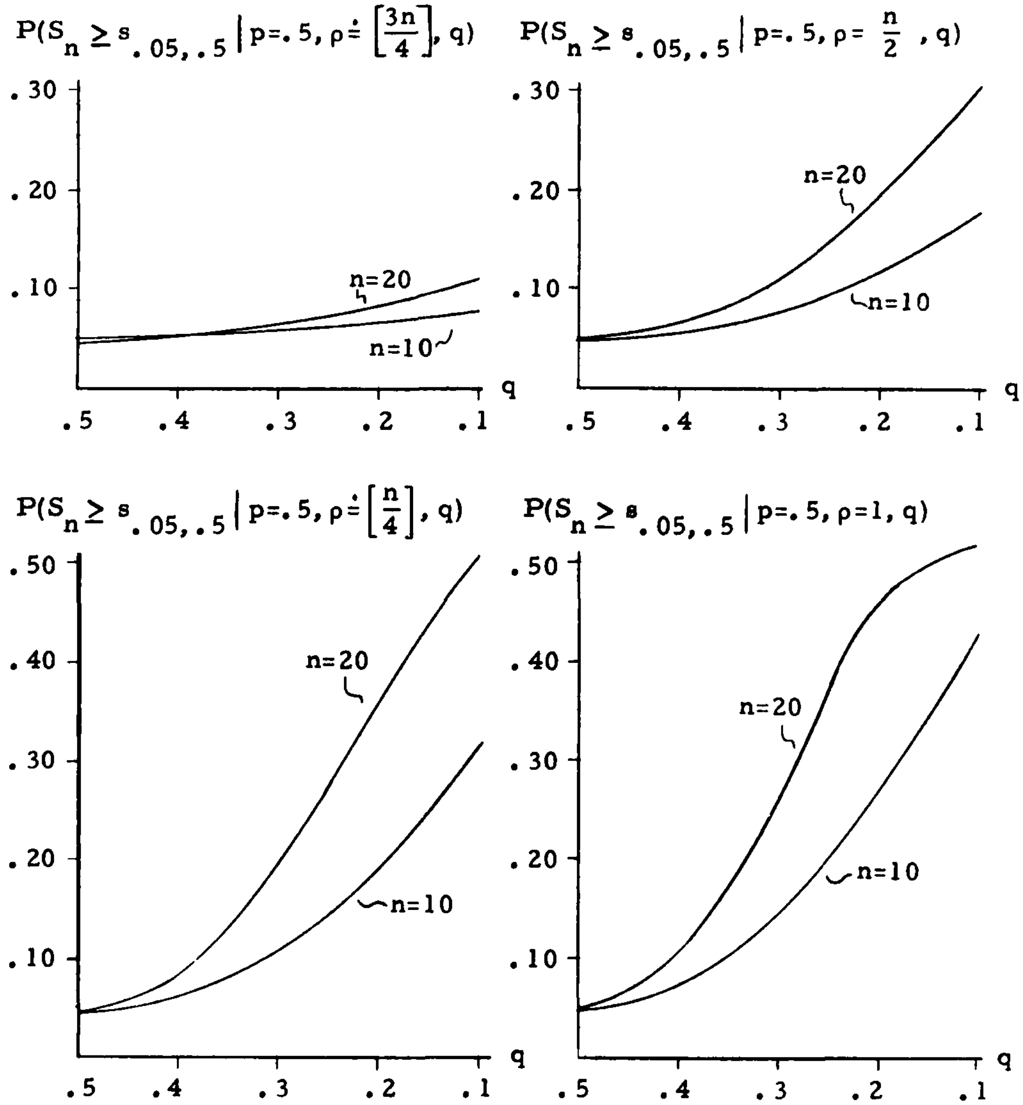


Figure 34.1 $P(S_n \geq s_{.05,.5} \mid p=.5, \rho, q)$.

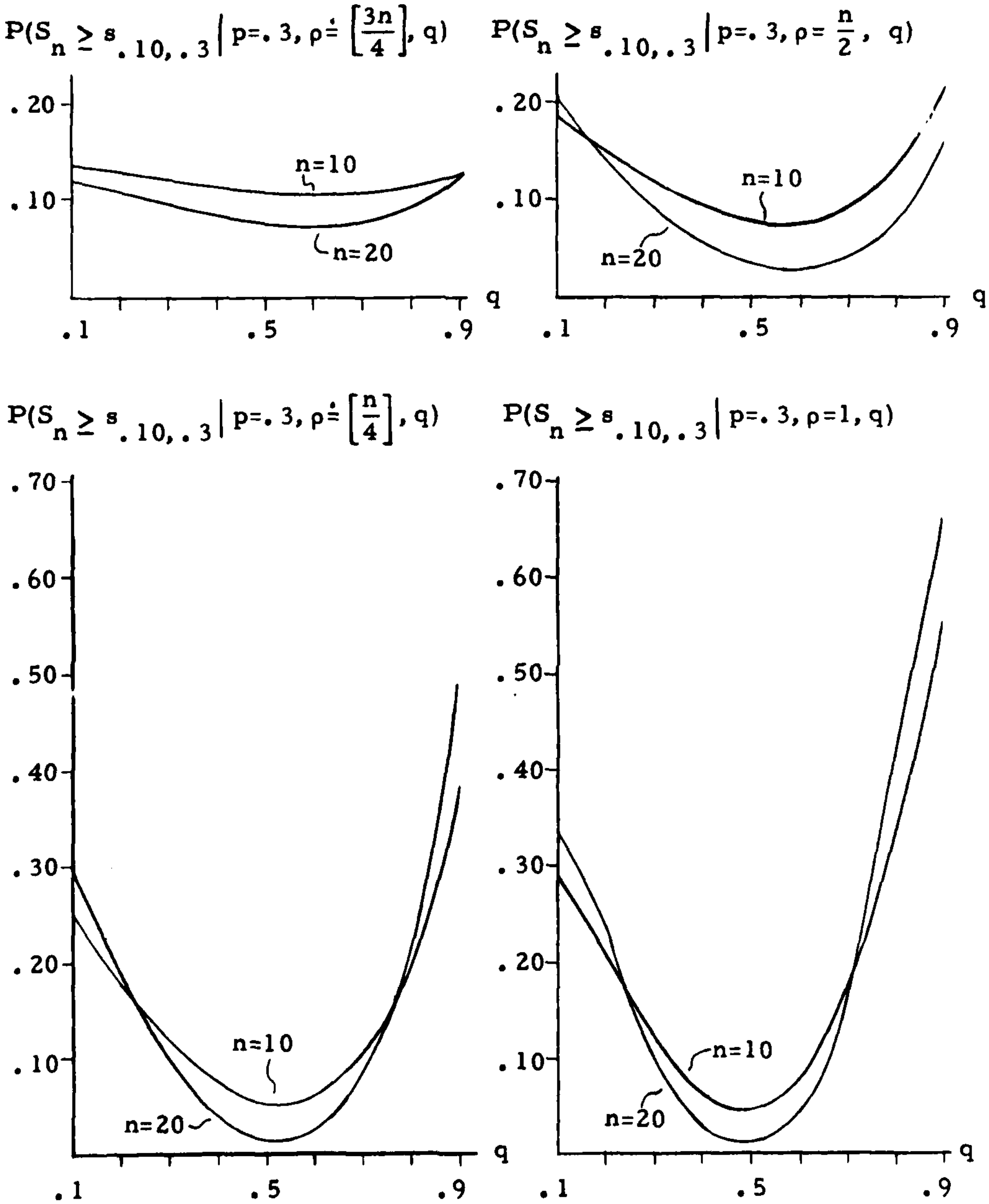


Figure 34.2 $P(S_n \geq s_{.10,.3} \mid p=.3, \rho, q)$.

function in q , $P(\hat{\rho}_{n,\alpha} = j \mid p = \frac{1}{2},\ \rho = j+i, q)$ in general decrease as $q \longrightarrow p$ for $i > 0$ and increase as $q \longrightarrow p$ for $i < 0$.

The magnitude of $P(\hat{\rho}_{n,\alpha} = j \mid p = \frac{1}{2},\ \rho = j, q)$ relative to $\sum_i P(\hat{\rho}_{n,\alpha} = j \mid p = \frac{1}{2},\ \rho = j+i, q)$ tends to be larger for each j as $|p-q|$ increases; for no value of j is the relative magnitude largest however.

Table 35 lists $P(\hat{\rho}_{10,\alpha} = j \mid p = \frac{1}{2},\ \rho, q)$ for $\alpha = .018,\ .047,\ .127$, $j = 1, 5, 9$ and $q = .1, .3$; the probabilities are symmetric in q and 1-q. The relative magnitudes of $P(\hat{\rho}_{10,\alpha} = j \mid p = .5,\ \rho = j,\ q = .1)$ to $\sum_i P(\hat{\rho}_{10,\alpha} = j \mid p = .5,\ \rho = j \pm i,\ q = .1)$ for the same values of α and j are listed below.

9	.2205	.3321	.2304
5	.2534	.2748	.3045
1	.3190	.1784	.2754
j/α	.018	.047	.127

Since the two-sided test is unbiased for $p = \frac{1}{2}$ and since in application it is desirable to detect change points quickly, the same experimental design is suggested here as for the one-sided test.

3.4.4 The Estimation of q

As for the one-sided testing situation, when $\hat{\rho} > \rho$ the last $n-\hat{\rho}$ observations may be used to obtain an unbiased estimator of q .

Table 35. $P(\hat{\rho}_{10,\alpha} = j \mid p = \frac{1}{2}, \rho, q)$.

j = 1, α = .018

9	.1109	.1109
8	.1215	.1148
7	.1379	.1229
6	.1600	.1364
5	.1904	.1556
4	.2353	.1829
3	.3077	.2228
2	.4445	.2862
1	.8000	.4004
ρ/q	.1	.3

j = 5, α = .018

9	.1109	.1109
8	.1215	.1148
7	.1379	.1229
6	.1600	.1364
5	.1904	.1556
4	.0261	.0806
3	.0038	.0441
2	.0006	.0266
1	.0001	.0192
ρ/q	.1	.3

j = 9, α = .018

9	.1109	.1109
8	.0266	.0830
7	.0169	.0697
6	.0179	.0668
5	.0212	.0709
4	.0261	.0806
3	.0342	.0967
2	.0494	.1232
1	.0889	.1720
ρ/q	.1	.3

j = 1, α = .047

9	.3542	.3542
8	.4527	.3807
7	.4681	.3945
6	.4768	.4080
5	.4843	.4230
4	.4947	.4409
3	.5125	.4643
2	.5515	.4972
1	.8239	.5832
ρ/q	.1	.3

j = 5, α = .047

9	.0416	.0416
8	.0617	.0472
7	.0742	.0532
6	.0884	.0608
5	.1076	.0699
4	.0152	.0358
3	.0023	.0190
2	.0004	.0109
1	.0001	.0066
ρ/q	.1	.3

j = 9, α = .047

9	.3542	.3542
8	.1153	.2899
7	.0778	.2566
6	.0690	.2370
5	.0703	.2295
4	.0753	.2314
3	.0849	.2414
2	.1059	.2603
1	.1137	.2691
ρ/q	.1	.3

j = 1, α = .127

9	.1308	.1308
8	.1439	.1347
7	.1668	.1438
6	.2003	.1585
5	.2498	.1785
4	.3374	.2085
3	.3929	.2388
2	.4326	.2760
1	.7809	.3884
ρ/q	.1	.3

j = 5, α = .127

9	.1077	.1077
8	.1067	.1074
7	.1155	.1105
6	.1610	.1285
5	.2405	.1584
4	.0450	.0927
3	.0078	.0558
2	.0041	.0438
1	.0015	.0378
ρ/q	.1	.3

j = 9, α = .127

9	.1308	.1308
8	.0366	.1026
7	.0277	.0935
6	.0290	.0920
5	.0362	.0969
4	.0514	.1049
3	.0651	.1242
2	.0830	.1445
1	.1078	.1792
ρ/q	.1	.3

CHAPTER 4

CONCLUSION

In Section 1 properties of the one-sided test statistic S_n^+, the generalized likelihood ratio test statistic $C_{p,q}$ and Page's test statistic M are reviewed; circumstances indicating preference for use of one of these statistics over the others are outlined. Suggestions for experimental design using the test statistics S_n and S_n^+ are presented in Section 2. Lastly directions for future research are discussed in Section 3.

4.1 S_n^+ or $C_{p,q}$ or M

Advantages and appropriateness for use of one of the test statistics over the others are discussed.

Of the three test statistics, construction of critical regions and evaluation of the statistic for a sample sequence are more difficult for $C_{p,q}$. Furthermore q, as well as p, must be known to perform the required calculations; in general it is not reasonable to expect that q is known however. Tabulation of critical regions for this test take up a great deal of space. Less effort is required to construct critical regions and evaluate the statistic for a sample sequence for the test statistics S_n^+ and M; only knowledge of p is required.

The types of sequences in the critical regions of the tests indicate the appropriateness of the statistics for the hypothesis being tested. The tendency for the tests based on $C_{p,q}$ and M is to put those sequences in the critical region for which the estimate of the change point is 0 or near 0. Hence detection of later changes is difficult or impossible for many values of p; for the one-sided tests with $p > q$ this tendency becomes more severe as p decreases. It appears that the tests based on $C_{p,q}$ and M more appropriately are testing no change versus a change either <u>before or after</u> sampling began. On the other hand, for S_n^+ whenever the null hypothesis of no change is rejected, each point $p \in \{1, \cdots, n-1\}$ has positive probability of being estimated regardless of the value of p; moreover for any reasonable size of test and for any value of p, the change point cannot be estimated to have occurred before sampling began. The test based on S_n^+ is more appropriately testing no change versus a change <u>after</u> sampling began.

For $C_{p,q}$ and M, rejection of the null hypothesis is difficult when p is small; for example when $n = 5, 10, 20$ and $p = .1$ the smallest values of α for which the null hypothesis can be rejected are $\alpha = .59, .35, .12$, respectively. When p is large, however, this is not a problem; for $n = 5$, $p = .9$ and $q = .7$, the critical regions contain 26 of the 32 sequences for $\alpha = .08$. For S_n^+, $P(S_n^+ \geq s^+ | p) = P(S_n^+ \geq s^+ | 1-p)$; the rejection of the null hypothesis can be realized for a much larger interval of p values.

A disturbing property of Page's test is that $x_1, \cdots, x_n$ and $x_n, \cdots, x_1$ give the same value of M. His test is behaving more like a two-sided test than a one-sided test as is proposed.

A comparison of the power functions for the tests based on S_n^+, $C_{p,q}$ and M indicates that the power of S_n^+ is larger than the power of $C_{p,q}$ or M when ρ is large and $p > \frac{1}{2}$, or p is small, and is competitive with the power of $C_{p,q}$ or M when ρ is large and $p \in (\delta, \frac{1}{2}]$ where δ is not small; the powers for $C_{p,q}$ and M are larger than that for S_n^+ otherwise.

In summary, the test based on $C_{p,q}$ or M is preferred if one suspects that a change is more likely to have occurred either before sampling began or shortly thereafter and if p is not small. If p is small or if one suspects that a change is more likely to have occurred later in the sequence, the test based on S_n^+ is preferred; hence for those situations where detecting a change quickly is of importance, the test based on S_n^+ is more appropriate.

4.2 Suggestions for Experimental Design

For the general problem where $Y_i \sim F(y; \theta)$ for $i = 1, \cdots, \rho$ and $Y_i \sim F(y; \theta')$ for $i = \rho+1, \cdots, n$ and θ is a location parameter, the proposed change point tests can be used to construct the appropriate one- or two-sided nonparametric tests for detecting a change in the model. The construction requires definition of the X_i's using some function of the location parameter, say $g(\theta)$, such that

$$X_i = \begin{cases} 1 & \text{if} \quad Y_i \in \{g(\theta):\ \theta \in \Phi_0\} \\ -1 & \text{if} \quad Y_i \notin \{g(\theta):\ \theta \in \Phi_0\} \end{cases} ,$$

$i = 1, \cdots, n$; the function $g(\theta)$ determines the probability $p = P(X_i = 1 \mid H_o)$. The choice of the function $g(\cdot)$ should be made to optimize properties of the test. Since properties of the one-sided test based on S_n^+ are better than those for the two-sided test based on S_n, the one-sided test is preferred; if one has reason to suspect a change in the location parameter in one direction, this information should be used to construct a one-sided test. Depending on the situation one may also have information as to where an anticipated change point might have occurred. This information should also be used advantageously.

Since detection of change points quickly is generally of concern, it is suggested that an anticipated change point be placed near the end of the sequence of observations $Y_1, \cdots, Y_n$. When the distribution F and the alternative θ' is not known except that it is say less than θ, choosing $g(\theta)$ to be the sample median of the distribution $F(x; \theta)$ is appropriate; past data are used to estimate the median.

4.3 Directions for Future Research

The manners in which the test statistics S_n^+ and S_n may be modified or extended to improve on this intuitive test for a change point are many.

For those instances in which a change in the model may have occurred before sampling began, one could modify the statistic S_n^+ as follows:

$$S_n^{+*} = \sup \left[1 - \frac{1}{n} \sum_{i=1}^{n} X_i, \sup_{r=1,\cdots,n} \left\{ \frac{1}{r} \sum_{i=1}^{r} X_i - \frac{1}{n-r} \sum_{i=r+1}^{n} X_i \right\} \right].$$

A similar modification may be made for S_n. There is good reason to suspect that the power of the test based on this statistic will be greater than that for S_n^+ when ρ is small. Alternatively when a change in the model may have occurred before sampling began, one might construct a test based on the union-intersection principle using S_n^+ and Page's test statistic M.

Attempts for finding approximate, rather than exact, critical regions and properties of the tests could be pursued to obtain a change point test for larger values of n. The behavior of the test statistics for $n \leq 20$ suggest that such attempts are possible.

The situation where the X_i's are not distributed as Bernoulli random variables is also of interest.

The behavior of the test statistics when the observations are dependent could also be investigated.

REFERENCES

Chernoff, H. and Zacks, S. (1964), "Estimating the current mean of a normal distribution which is subjected to change in time," Annals of Mathematical Statistics, 35, 999-1018.

Hinkley, D. V. and Hinkley, E. A. (1970), "Inference about the change point in a sequence of binomial variables," Biometrika, 57, 477-88.

Page, E. S. (1955), "A test for a change in a parameter occurring at an unknown point," Biometrika, 42, 523-26.

DAT
FILM